Sojourn time approximations for a discriminatory-processor-sharing queue [1]

A. Izagirre[a,b,d,e], U. Ayesta[b,c,d,e], I.M. Verloop[a,e]
ane.izagirre@laas.fr, urtzi@laas.fr, maaike.verloop@enseeiht.fr

[a]CNRS ; IRIT ; 2 rue C. Camichel, F-31071 Toulouse, France
[b]CNRS ; LAAS ; 7 avenue du colonel Roche, F-31400 Toulouse, France
[c]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[d]UPV/EHU, Univ. of the Basque Country, 20018 Donostia, Spain
[e]Université de Toulouse ; INP, INSA ; *IRIT*, *LAAS* ; F-31400 Toulouse, France

We study a multi-class time-sharing discipline with relative priorities known as Discriminatory Processor Sharing (DPS), which provides a natural framework to model service differentiation in systems. The analysis of DPS is extremely challenging and analytical results are scarce. We develop closed-form approximations for the mean conditional (on the service requirement) and unconditional sojourn times. The main benefits of the approximations lie in its simplicity, the fact that it applies for general service requirements with finite second moments, and that it provides insights into the dependency of the performance on the system parameters. We show that the approximation for the mean conditional and unconditional sojourn time of a customer is decreasing as its relative priority increases. We also show that the approximation is exact in various scenarios, and that it is uniformly bounded in the second moments of the service requirements. Finally we numerically illustrate that the approximation for exponential, hyperexponential and Pareto service requirements is accurate across a broad range of parameters.

Categories and Subject Descriptors: D.4.8 [**Operating systems**]: Performance—*Queueing theory*

General Terms: Performance

Additional Key Words and Phrases: Discriminatory-processor-sharing, sojourn time, light-traffic, heavy-traffic, interpolation

1. INTRODUCTION

The Discriminatory Processor Sharing queue (DPS) is a versatile queueing model providing a natural framework to model service differentiation in systems. It is a multi-class extension of the well-studied

[1]A conference version of this paper was published in [Izagirre et al. 2014].

DOI 10.1145/2812807 http://doi.acm.org/10.1145/2812807

egalitarian Processor Sharing (PS) policy, where the various classes are assigned positive weight factors. The service capacity is shared simultaneously among all customers present in proportion to the respective class-dependent weights. More precisely, given there are K classes of customers, if at time t there are $n_k(t)$ class-k customers present in the system, $k = 1, \ldots, K$, under DPS each class-k customer is served at rate $g_k / \sum_{j=1}^{K} g_j n_j(t)$, where $g_1, \ldots, g_K$, are the class-dependent weights. The DPS queue has received lot of attention due to its application to model the impact of service differentiation in systems.

When all the weights are equal, the DPS queue is equivalent to the PS queue. The PS queue has gained a prominent role in evaluating the performance of a variety of resource allocation mechanisms (see for example [Kleinrock 1976; Kelly 1979; Yashkov 1987]), and in recent years it has received renewed attention as a convenient abstraction for modeling the flow-level performance of bandwidth-sharing protocols in packet-switched networks, in particular TCP, see for example [Fredj et al. 2001; Roberts 2004]. In multiple practical situations, the actual service shares that users obtain may show substantial variation among users with heterogeneous characteristics. For example, TCP flows that share a common bottleneck link but traverse distinct routes, may experience diverse packet loss rates and round-trip delays. Besides TCP-related effects, the heterogeneity in bandwidth shares may also be due to deliberate service differentiation among competing flows (for example different quality-of-service in the Internet). For instance packet scheduling algorithms, such as Weighted Fair Queueing (WFQ) and Weighted Round-Robin (WRR), have been proposed as potential instruments to implement differentiated bandwidth sharing.

In this context, the *Discriminatory* Processor-Sharing (DPS) provides a natural approach for modeling the flow-level performance of TCP. The DPS model was introduced in [Kleinrock 1967]. Despite the simplicity of the model description and the fact that the properties of the egalitarian PS queue are quite thoroughly understood, the analysis of DPS has proven to be extremely difficult. For example, results on an important basic metric like the mean sojourn time in the system have only been derived in a very implicit manner or under certain limiting regimes (time-scale decomposition, heavy-traffic, overload etc.).

In the seminal paper [Fayolle et al. 1980] the authors studied the mean conditional (on the service requirement) and unconditional sojourn time. For general service time distributions, the authors obtained the mean conditional sojourn time as the solution of a system of integro-differential equations. In addition, the authors provided a thorough analysis for the case of exponentially distributed service requirements. However, except for the case of two classes, no closed-form expression is available and numerical analysis is needed in order to calculate the mean sojourn times. Since we use the results of [Fayolle et al. 1980] in order to evaluate the accuracy of our approximation, we will give further details on them in Section 2. [Avrachenkov et al. 2005] established that the mean queue lengths of all classes are finite under the usual stability condition, regardless of the higher-order moments of the service requirements. Asymptotics of the sojourn time have also received considerable attention for example in [Borst et al. 2005; Borst et al. 2006]. An important result in this area establishes the asymptotic equivalence between the sojourn time distribution and the service time distribution. Time-scale separations have been studied in [van Kessel et al. 2005] and [Boxma et al. 2006]. In particular, the authors of [Boxma et al. 2006] approximate the distribution of the sojourn time for a DPS queue with admission control. However the expressions derived in [Boxma et al. 2006] need to be solved numerically. The performance of DPS in overload and its application to model TCP flows is considered in [Altman et al. 2004]. The application of DPS to analyse the performance of TCP is also considered in [Kherani and Núñez-Queija 2006]. For more applications of DPS in communication networks see [Bu and Towsley 2001; Cheung et al. 2005; Hayel and Tuffin 2005]. DPS under a heavy-traffic regime (when the traffic load approaches the available capacity) was analysed in [Grishechkin 1992]

assuming finite second moments of the service requirement distributions. Subsequently, assuming exponential service requirement distributions, a direct approach to establish a heavy-traffic limit for the joint queue length distribution was described in [Rege and Sengupta 1996] and extended to *phase-type* distributions in [Verloop et al. 2011]. We refer to Section 2 for more details on heavy-traffic results. In [Izagirre et al. 2015] an interpolation approximation is derived for the steady-state distribution of the queue length and waiting time of DPS and a relative priorities system. Game-theoretic aspects of DPS have been studied in [Wu et al. 2012] and [Hassin and Haviv 2003]. For an extensive overview of the literature on DPS we refer to the survey [Altman et al. 2006].

Motivated by the difficulty in analyzing the system in exact form, in this paper we derive a closed-form approximation for the mean conditional and unconditional sojourn time in the system. We first obtain a light-traffic approximation using the framework obtained in [Reiman and Simon 1989]. To the best of our knowledge, we are the first to obtain a light-traffic approximation of a time-sharing system, that is, when all users in the system simultaneously get served. We then use results from the heavy-traffic literature in order to obtain a polynomial approximation for any value of the load of the mean conditional sojourn time for service requirements with finite second moments. Unconditioning on the service time distribution, this allows us to readily obtain an approximation for the mean unconditional sojourn time. We will show that in some cases our approximation becomes exact, for example when there is only one class in the system or when all the weights are the same. The approximation provides insights into the performance of the system. We show that the approximation for the mean conditional sojourn time of a class-k user is decreasing (resp. increasing) as the weight g_k (resp. g_j, $j \neq k$) increases. Another important observation is that the approximation is uniformly bounded in the second moments of the service requirements. This was a major property of PS, which is in sheer contrast with FCFS queues, where the mean waiting time explodes as the second moment grows. In the particular case of exponential service time distributions we see that the expression greatly simplifies and that the approximation for the mean unconditional sojourn time is exact when the mean service times for all classes are the same. Finally, we numerically investigate the accuracy of the approximation by comparing it with the exact results obtained in [Fayolle et al. 1980]. We consider exponential, hyper-exponential and Pareto service time distributions, and our results show that our approximation works extremely well across various parameter values. An important benefit of the approximation is that it provides insights into the dependency of the performance on the system parameters (weights, service time distributions, etc), and we thus believe it will provide an interesting tool in order to implement service-differentiation in real systems.

The remainder of the paper is organized as follows. In Section 2 we provide a detailed model description and gather results from [Fayolle et al. 1980] and [Grishechkin 1992] that will be used in the paper. In Section 3 we develop a light-traffic analysis. The light and heavy-traffic interpolation approximation for the conditional and unconditional sojourn time is presented in Section 4. Section 5 presents the results for the particular case of exponentially distributed service requirements. In Section 6 we numerically test the accuracy of the obtained approximations.

2. MODEL DESCRIPTION AND PRELIMINARIES

We consider a multi-class single-server queue with K classes of customers. Class-k customers, $k = 1, \ldots, K$, arrive according to independent Poisson processes with rate $\lambda_k \geq 0$. We denote the overall arrival rate by $\lambda = \sum_{k=1}^{K} \lambda_k$. A class-$k$ customer has i.i.d generally distributed service requirements B_k and we assume $\mathbb{E}[B_k^2] < \infty$, $k = 1, \ldots, K$. The traffic intensity for class-k customers is denoted by $\rho_k := \lambda_k \mathbb{E}[B_k]$ and the total traffic intensity is denoted by $\rho := \sum_{k=1}^{K} \rho_k = \sum_{k=1}^{K} \lambda_k \mathbb{E}[B_k] = \lambda \sum_{k=1}^{K} \alpha_k \mathbb{E}[B_k] = \lambda \mathbb{E}[B]$, where $\alpha_k = \lambda_k / \lambda$ denotes the probability that an arrival is of class k and the random variable B is the service requirement of an arbitrary arriving customer.

The K customer classes share a common resource of capacity one. There are strictly positive weights $g_1, \ldots, g_K$ associated with each of the classes. Whenever there are n_k class-k customers, $k = 1, \ldots, K$, in the system, each class-k customer is served at rate

$$\frac{g_k}{\sum_{j=1}^{K} n_j g_j}.$$

We denote by $S_k(\lambda, b)$ the conditional sojourn time of a tagged class-k customer with a given service requirement b, when the arrival rate is λ. We are interested in approximating $\overline{S}_k(\lambda, b) := \mathbb{E}[S_k(\lambda, b)]$, the mean conditional sojourn time of the tagged class-k customer. We further denote by $\overline{S}_k(\lambda) := \int_0^\infty \overline{S}_k(\lambda, b) \mathrm{d}F_k(b)$ the mean unconditional sojourn time of the tagged class-k customer, where $\mathbb{P}(B_k \leq b) = F_k(b)$ is the distribution function of B_k, and $\overline{S}(\lambda) := \sum_{k=1}^{K} \alpha_k \overline{S}_k(\lambda)$ is the mean unconditional sojourn time of *an arbitrary customer*.

The analysis of DPS is extremely difficult compared to that of egalitarian PS, which arises as a special case when all g_k are equal. In [Fayolle et al. 1980] the authors obtained that the derivatives of the mean conditional sojourn times of the various classes satisfy the following system of integro-differential equations:

$$\frac{\partial \overline{S}_k(\lambda, b)}{\partial b} = 1 + \lambda \sum_{j=1}^{K} \int_0^\infty \alpha_j \frac{g_j}{g_k} \frac{\partial \overline{S}_j(\lambda, y)}{\partial y} [1 - F_j(y + \frac{g_j}{g_k} b)] \mathrm{d}y + \lambda \int_0^b \frac{\partial \overline{S}_k(\lambda, y)}{\partial y} \sum_{j=1}^{K} \alpha_j \frac{g_j}{g_k} [1 - F_j(\frac{g_j}{g_k}(b - y))] \mathrm{d}y, \quad (1)$$

for $k = 1, \ldots, K$. The natural boundary conditions are $\overline{S}_k(\lambda, 0) = 0$, $k = 1, \ldots, K$.

The only known analytical solution for this system of equations has been obtained under the assumption of exponentially distributed service requirements. In this case we denote by $\mu_j := 1/\mathbb{E}[B_j]$, $\forall j$. In [Fayolle et al. 1980] it is proved that

$$\overline{S}_k(\lambda, b) = \frac{b}{1 - \rho} + \sum_{j=1}^{m} \frac{g_k c_j \beta_j + d_j}{\beta_j^2} \left(1 - \mathrm{e}^{-\beta_j b / g_k}\right), \quad (2)$$

where $-\beta_j$, $j = 1, 2, \ldots, m$, are the m distinct negative roots of

$$\sum_{j=1}^{K} \frac{\lambda_j g_j}{\mu_j g_j + s} = 1, \quad (3)$$

and where c_j and d_j, $j = 1, \ldots, m$, are a function of the input parameters and β_j, $j = 1, \ldots, m$.

Furthermore, for the mean unconditional sojourn time with exponentially distributed service requirements, it is shown in [Fayolle et al. 1980] that $\overline{S}_k(\lambda)$, $k = 1, \ldots, K$, is the unique solution of the following system of equations:

$$\overline{S}_k(\lambda) \left(1 - \sum_{j=1}^{K} \frac{\lambda_j g_j}{\mu_j g_j + \mu_k g_k}\right) - \sum_{j=1}^{K} \frac{\lambda_j g_j \overline{S}_j(\lambda)}{\mu_j g_j + \mu_k g_k} = \frac{1}{\mu_k}. \quad (4)$$

A closed-form solution for this system of equations (4) is available only for the case of $K = 2$, and is given by

$$\overline{S}_1(\lambda) = \frac{1}{\mu_1(1 - \rho)} \left(1 + \frac{\mu_1 \rho_2 (g_2 - g_1)}{D}\right) \quad (5)$$

and

$$\overline{S}_2(\lambda) = \frac{1}{\mu_2(1-\rho)}\left(1 + \frac{\mu_2\rho_1(g_1-g_2)}{D}\right), \tag{6}$$

where $D = \mu_1 g_1(1-\rho_1) + \mu_2 g_2(1-\rho_2)$.

The above shows how hard and challenging it is to study analytically the DPS model. For this reason, as mentioned in the introduction, research has focused on analysing the DPS queue in limiting regimes, like tail asymptotics, heavy-traffic limits, fluid limits etc. In this paper, we take a different approach, and we develop a light and heavy-traffic interpolation based approximation for $\overline{S}_k(\lambda, b)$ and $\overline{S}_k(\lambda)$. In the numerical section we will use Equations (2), (4)-(6) in order to numerically verify the accuracy of our interpolation approximations.

The approximation is obtained by interpolating the mean sojourn times obtained under the light-traffic regime and the heavy-traffic regime.

The light-traffic regime consists in letting $\rho \downarrow 0$, or equivalently $\lambda \downarrow 0$. Hence, it concerns the performance when the system is almost empty. Therefore, in Section 3 we analyze the mean conditional sojourn time in the light-traffic regime.

The heavy-traffic regime consists in letting $\rho \uparrow 1$, or equivalently $\lambda \uparrow 1/\mathbb{E}[B]$. Hence, it concerns the performance when the system is close to congestion. Heavy-traffic results have been obtained in [Grishechkin 1992; Rege and Sengupta 1996; Verloop et al. 2011]. For our analysis, we use the results by Grishechkin [Grishechkin 1992, Theorem 4.1] who studied a general processor-sharing system of which our model is a particular case. In particular, for the DPS queue Grishechkin derives the distribution of the conditional sojourn times, scaled by $1-\lambda\mathbb{E}[B] = 1-\rho$, as $\lambda \uparrow 1/\mathbb{E}[B]$. In particular, the mean of this distribution is given by

$$\mathbb{E}[\lim_{\lambda\uparrow 1/\mathbb{E}[B]}(1-\lambda\mathbb{E}[B])S_k(\lambda, b)] = \frac{b}{g_k}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j \mathbb{E}[B_j^2]/g_j}. \tag{7}$$

For our interpolation result, we are interested in $\lim_{\lambda\uparrow 1/\mathbb{E}[B]}(1-\lambda\mathbb{E}[B])\overline{S}_k(\lambda, b) = \lim_{\lambda\uparrow 1/\mathbb{E}[B]}(1-\lambda\mathbb{E}[B])\mathbb{E}[S_k(\lambda, b)]$. Although we cannot verify that the limit and expectation can be interchanged, we use the expression in (7) as an approximation for $\lim_{\lambda\uparrow 1/\mathbb{E}[B]}(1-\lambda\mathbb{E}[B])\overline{S}_k(\lambda, b)$. Numerical experiments as performed in [Verloop et al. 2011] indicate that indeed the limits can be interchanged.

3. LIGHT-TRAFFIC ANALYSIS

In this section we analyse the mean conditional sojourn time of the tagged class-k customer under the light-traffic regime. The light-traffic regime concerns the performance of the system for small values of the arrival rate λ, i.e., when the system is almost empty. We will approximate $\overline{S}_k(\lambda, b)$ by a Taylor series expansion of $\overline{S}_k(\lambda, b)$ at $\lambda = 0$. Assuming that the first n derivatives of $\overline{S}_k(\lambda, b)$ with respect to λ at $\lambda = 0$ exist we have the following approximation for the mean conditional sojourn time of a class-k customer when λ is close to zero:

$$\overline{S}_k^{LT}(\lambda, b) := \overline{S}_k^{(0)}(0, b) + \lambda\overline{S}_k^{(1)}(0, b) + \cdots + \frac{\lambda^n}{n!}\overline{S}_k^{(n)}(0, b). \tag{8}$$

We will refer to this as the light-traffic approximation of order n. Here $\overline{S}_k^{(0)}(0, b) = \overline{S}_k(0, b)$ and we refer to it as the *zeroth* light-traffic derivative. Moreover, $\overline{S}_k^{(m)}(0, b), m = 1, 2, \ldots,$ denotes the m-th derivative of $\overline{S}_k(\lambda, b)$ with respect to λ at $\lambda = 0$, i.e., $\frac{\partial^m \overline{S}_k(\lambda,b)}{\partial\lambda^m}\Big|_{\lambda=0}$. We have based our analysis on Reiman and Simon [Reiman and Simon 1989] where it is shown how to obtain the derivatives of arbitrary order $m \geq 0$ at $\lambda = 0$ under a general admissibility condition. Following the discussion in [Reiman and Simon

1989, Appendix A] we make the next assumption on the service requirements B_k

$$\mathbb{E}[\mathrm{e}^{\eta B_k}] = \sum_{n=0}^{\infty} \frac{\eta^n}{n!} \mathbb{E}[B_k^n] < \infty \tag{9}$$

for some $\eta > 0, \forall k$. This finite exponential moment condition requires that all moments of the service requirement B_k to be finite. Equation (9) entails admissibility; it is likely stronger than needed but its purpose here is to provide a convenient framework where calculations can be justified.

In this paper we set $n = 1$ in (8) as this will already provide us with an accurate approximation of the performance. Let $A(s,t)$ denote the number of arrivals in the interval $[s,t)$ in addition to the tagged customer who is assumed to arrive at time 0. Then, the zeroth and first light-traffic derivatives satisfy

$$\overline{S}_k(0,b) := \mathbb{E}\left[S_k(0,b) \middle| A(-\infty,\infty) = 0\right] \tag{10}$$

and

$$\overline{S}_k^{(1)}(0,b) := \int_{-\infty}^{\infty} \left(\mathbb{E}\left[S_k(0,b) \middle| A(-\infty,\infty) = 1, \tau = t\right] - \mathbb{E}\left[S_k(0,b) \middle| A(-\infty,\infty) = 0\right] \right) dt, \tag{11}$$

where τ is the arrival time of the first customer, see [Reiman and Simon 1989]. In Appendix A we provide a brief intuitive approach of how to obtain the light-traffic derivatives (10) and (11).

3.1 Light-traffic approximation

In this section we derive the first-order light-traffic approximation.

Equation (10) represents the situation where nobody enters the system except the tagged customer. Therefore, $\overline{S}_k(0,b)$ is equal to the service requirement of the tagged customer, which we denote by b. Hence,

$$\overline{S}_k(0,b) = b. \tag{12}$$

Let us denote by $S_{k,t,u_t,b_{u_t}}$ the sojourn time of the tagged class-k customer when there is exactly one arrival at time t on $\mathbb{R}$, u_t describing the class of the customer arriving at time t and b_{u_t} denoting the service requirement of the customer arriving at time t. Hence, $\mathbb{E}\left[S_k(0,b) \middle| A(-\infty,\infty) = 1, \tau = t\right]$ and $\mathbb{E}[S_{k,t,U_t,B_{U_t}}]$ are equivalent, where U_t and B_{U_t} are dependent random variables and are distributed as follows: with probability α_i we have $U_t = i$ and B_{U_t} is distributed as B_i, $i = 1,\ldots,K$. We can write $S_{k,t,u_t,b_{u_t}}$ as follows:

$$S_{k,t,u_t,b_{u_t}} = \begin{cases} t + b_{u_t} + b & \text{if } t \le 0 \le t + b_{u_t} \text{ and } \frac{b}{g_k} > \frac{t+b_{u_t}}{g_{u_t}} \\ \frac{g_k+g_{u_t}}{g_k} b & \text{if } t \le 0 \le t + b_{u_t} \text{ and } \frac{b}{g_k} \le \frac{t+b_{u_t}}{g_{u_t}} \\ b & \text{if } t + b_{u_t} < 0 \\ b + b_{u_t} & \text{if } 0 < t < b \text{ and } \frac{b-t}{g_k} > \frac{b_{u_t}}{g_{u_t}} \\ t + (b-t)\frac{g_k+g_{u_t}}{g_k} & \text{if } 0 < t < b \text{ and } \frac{b-t}{g_k} \le \frac{b_{u_t}}{g_{u_t}} \\ b & \text{if } 0 < b < t. \end{cases} \tag{13}$$

The first expression describes the case where the customer arrives before the tagged customer and leaves after the tagged customer arrives, but before the tagged customer leaves. Hence, by the work conserving property, the tagged customer stays in the system until all the work present at time 0 is done, that is, $b_{u_t} - (-t) + b$. We recall that the work-conserving property states that as long as the

system is non-empty, the server does not idle. The second term describes the case where the other customer is in the system at time 0 and is still present as the tagged customer departs. Hence, the tagged class-k customer is served at rate $\frac{g_k}{g_k+g_{u_t}}$, so that its sojourn time is $b\left(\frac{g_k}{g_k+g_{u_t}}\right)^{-1}$. The fourth expression describes the case where the customer arrives after the tagged customer and leaves before the tagged customer. Hence, by the work-conserving property of the system, the sojourn time of the tagged class-k customer is given by the total amount of work that needs to be done, that is, $b+b_{u_t}$. The fifth term describes the case where the customer arrives after the tagged customer, and departs after the tagged customer departs. Then, the sojourn time of the tagged customer is composed of t, the time it was in the system until the customer arrived, plus $(b-t)\left(\frac{g_k}{g_k+g_{u_t}}\right)^{-1}$, the remaining service requirement multiplied by the inverse of the rate at which the the tagged class-k customer is served. The third and sixth case is when the tagged customer does not coincide with the other customer. Hence, the sojourn time is given by its service requirement, b.

From Equations (11), (12) and (13) we then obtain the following expression for the first derivative.

LEMMA 3.1. *We have*

$$\overline{S}_k^{(1)}(0,b) = \int_{\mathbb{R}} \left(\mathbb{E}\left[S_k(0,b)\middle|A(-\infty,\infty)=1,\tau=t\right]-b\right)\mathrm{d}t = \int_{\mathbb{R}} (\mathbb{E}[S_{k,t,U_t,B_{U_t}}]-b)\mathrm{d}t$$

$$= \mathbb{E}\left[\frac{1}{2}\big(1+\frac{g_k}{g_{U_t}}\big)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 - \big(b\frac{g_{U_t}}{g_k}+\frac{g_k}{g_{U_t}}B_{U_t}\big)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} + \frac{g_k+g_{U_t}}{g_k}bB_{U_t}\right]. \quad (14)$$

See Appendix B for the proof.

From (8), (12) and (14) we now derive the following approximation for the mean conditional sojourn time when λ is small.

COROLLARY 3.2. *The light-traffic approximation (of order 1) of the mean conditional sojourn time for a tagged class-k customer with service requirement b is given by*

$$\overline{S}_k^{LT}(\lambda,b) = \overline{S}_k(0,b) + \lambda\overline{S}_k^{(1)}(0,b)$$

$$= b(1+\rho) + \lambda\mathbb{E}\left[\frac{1}{2}\big(1+\frac{g_k}{g_{U_t}}\big)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 - \big(b\frac{g_{U_t}}{g_k}+\frac{g_k}{g_{U_t}}B_{U_t}\big)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} + \frac{g_{U_t}}{g_k}bB_{U_t}\right]. \quad (15)$$

REMARK 1. *We consider in the paper the light-traffic approximation of order 1. Calculating the second light-traffic derivative would imply having to consider events that either 0, 1, or 2 customers arrive in the system (besides the tagged customer). The latter would result in going through 22 different cases, while for the first derivate we only needed to go through 6 cases, see Equation (13).*

We will see in Sections 4, 5 and 6 that already the first order light-traffic approximation provides an insightful and accurate approximation of the performance. We further refer to Appendix C where the light-traffic approximation of order 1 and of order 2 are numerically compared.

We can infer several nice properties from (15). For instance, we will show in Section 4 that (15) is decreasing in g_k and increasing in g_j, $j \neq k$. In other words, the approximation for the mean sojourn time reduces as its own weight increases, and it increases as the weight of any other class increases. Another interesting observation is that the light-traffic approximation of the mean conditional sojourn time can be uniformly bounded in the second moment. This important feature helps obtaining a good

performance in the presence of highly variable service distributions (like the ones observed in nowadays communication systems). See Section 4.4 for details.

4. LIGHT AND HEAVY-TRAFFIC INTERPOLATION

In this section we present the light and heavy-traffic interpolation result. This technique was popularized by Reiman and Simon [Reiman and Simon 1988a; 1988b; 1989] and consists in interpolating

$$t_k(\lambda) := (1-\rho)\overline{S}_k(\lambda,b) = (1-\lambda\mathbb{E}[B])\overline{S}_k(\lambda,b),$$

by a polynomial $\hat{t}_k(\lambda)$ of order $n+1$:

$$\hat{t}_k(\lambda) = h_0 + h_1\lambda + \ldots + h_{n+1}\lambda^{n+1}. \tag{16}$$

To determine the coefficients $h_0,\ldots,h_n$ we use the so-called light-traffic conditions

$$\hat{t}_k(0) = t_k(0) \quad \text{and} \quad \hat{t}_k^{(m)}(0) = t_k^{(m)}(0), \text{ for } m = 1,\ldots,n, \tag{17}$$

and the heavy-traffic condition

$$\hat{t}_k\big((1/\mathbb{E}[B])^-\big) = t_k\big((1/\mathbb{E}[B])^-\big), \tag{18}$$

where $t_k\big((1/\mathbb{E}[B])^-\big)$ is given by $\dfrac{b}{g_k}\dfrac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j\mathbb{E}[B_j^2]/g_j}$, see (7).

Once we have obtained the coefficients we undo the normalisation so that

$$\overline{S}_k^{INT}(\lambda,b) := \frac{\hat{t}_k(\lambda)}{1-\lambda\mathbb{E}[B]}, \quad 0 \le \lambda < 1/\mathbb{E}[B], \tag{19}$$

provides an approximation for the mean conditional sojourn time $\overline{S}_k(\lambda,b)$. We refer to this approximation as the light and heavy-traffic interpolation of order $n+1$.

In the following proposition we characterise (19) in terms of the light-traffic derivatives and the heavy-traffic equation. To the best of our knowledge, this is a new result and it applies to any model, i.e., it is not restricted to the DPS model.

PROPOSITION 4.1. *The light and heavy-traffic interpolation of order $n+1$ can be written as*

$$\overline{S}_k^{INT}(\lambda,b) \;=\; \sum_{i=0}^{n}\frac{\lambda^i}{i!}\overline{S}_k^{(i)}(0,b) + t_k\big((1/\mathbb{E}[B])^-\big)\frac{(\lambda\mathbb{E}[B])^{n+1}}{1-\lambda\mathbb{E}[B]}. \tag{20}$$

PROOF. From the light-traffic condition (17) we obtain

$$h_0 = \overline{S}_k^{(0)}(0,b) \quad \text{and} \quad h_i = \frac{\overline{S}_k^{(i)}(0,b)}{i!} - \mathbb{E}[B]\frac{\overline{S}_k^{(i-1)}(0,b)}{(i-1)!}, i = 1,2,...,n,$$

and from the heavy-traffic condition (18) we obtain

$$\begin{aligned}
h_{n+1} \;&=\; \mathbb{E}[B]^{n+1}\left(t_k\big((1/\mathbb{E}[B])^-\big) - \sum_{i=0}^{n}\frac{h_i}{\mathbb{E}[B]^i}\right)\\
&=\; \mathbb{E}[B]^{n+1}\left(t_k\big((1/\mathbb{E}[B])^-\big) - \overline{S}_k^{(0)}(0,b) - \sum_{i=1}^{n}\frac{1}{\mathbb{E}[B]^i}\left(\frac{\overline{S}_k^{(i)}(0,b)}{i!} - \mathbb{E}[B]\frac{\overline{S}_k^{(i-1)}(0,b)}{(i-1)!}\right)\right)\\
&=\; \mathbb{E}[B]^{n+1}\left(t_k\big((1/\mathbb{E}[B])^-\big) - \frac{1}{\mathbb{E}[B]^n}\frac{\overline{S}_k^{(n)}(0,b)}{n!}\right).
\end{aligned}$$

Equation (20) follows after substituting these expressions in (19). □

REMARK 2. *In [Izagirre 2015, Chapter 2], Proposition 4.1 is generalized to the case in which the heavy-traffic scaling is different than* $1-\lambda\mathbb{E}[B]$.

Notice that in the previous section we derived the light-traffic derivatives up to order 1. Hence, this allows us to obtain the light and heavy-traffic interpolation of order 2 as stated in the following proposition.

PROPOSITION 4.2. *The light and heavy-traffic interpolation (of order 2) of the mean conditional sojourn time for a tagged class-*k *customer with service requirement* b *is given by*

$$\begin{aligned}&\overline{S}_k^{INT}(\lambda,b)\\&= b+\lambda b\mathbb{E}[B]+\lambda\mathbb{E}\left[\frac{1}{2}\big(1+\frac{g_k}{g_{U_t}}\big)\min\{B_{U_t},b\frac{g_{U_t}}{g_k}\}^2-\big(b\frac{g_{U_t}}{g_k}+\frac{g_k}{g_{U_t}}B_{U_t}\big)\min\{B_{U_t},b\frac{g_{U_t}}{g_k}\}+b\frac{g_{U_t}}{g_k}B_{U_t}\right]\\&\quad+\frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{b}{g_k}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^K\alpha_j\mathbb{E}[B_j^2]/g_j}.\end{aligned}\tag{21}$$

PROOF. This follows from Equation (20) together with (12), (14) and (7). □

In [Izagirre et al. 2014] Proposition 4.2 was proved by calculating the coefficients h_0,h_1,h_2 using Equations (17) and (18). Thanks to Proposition 4.1, the proof of Proposition 4.2 is now immediate.

In Section 6 we will numerically evaluate the accuracy of the approximation formulas derived in Proposition 4.2. In the subsections below, we first make several interesting observations.

REMARK 3. *In Appendix C we describe an alternative way to obtain the light-traffic derivatives. This method makes use of Equation (1), and hence it applies only to the DPS model. It allows us easily to derive higher order light-traffic approximations. We observe numerically, see Figures 14-17, that the light-traffic approximation (of order* 2*) gets more accurate, whereas the accuracy of the interpolation (of order* 3*) for intermediate loads does not necessarily get better.*

4.1 Processor Sharing

For the standard Processor Sharing queue the mean conditional sojourn time is known and is given by $b/(1-\rho)$, [Kelly 1997]. If either (i) there is only one class or (ii) all weights are the same, our model is equivalent to a processor-sharing queue. Below we will verify that our approximation as stated in (21) indeed coincides with $b/(1-\rho)$.

We first consider the case of one class, that is, $\alpha_i=0,\forall i\neq k$ and $\alpha_k=1$. Then Equation (21) is equal to

$$b(1+\rho)+\lambda_k\mathbb{E}\left[\min\{B_{U_t},b\}^2-\big(b+B_{U_t}\big)\min\{B_{U_t},b\}+bB_{U_t}\right]+b\frac{\rho^2}{(1-\rho)}=b(1+\rho+\frac{\rho^2}{(1-\rho)})=\frac{b}{1-\rho},$$

where we used that $\min\{B_{U_t},b\}^2-\big(b+B_{U_t}\big)\min\{B_{U_t},b\}+bB_{U_t}=0$.

We now assume all weights are the same, i.e., $g_i=g_k,\forall i,k=1,\ldots,K$. Equation (21) is then equal to

$$\begin{aligned}&b(1+\rho)+\lambda\mathbb{E}\left[\min\{B_{U_t},b\}^2-\big(b+B_{U_t}\big)\min\{B_{U_t},b\}+bB_{U_t}\right]+\frac{b\rho^2}{(1-\rho)}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^K\alpha_j\mathbb{E}[B_j^2]}\\&=b(1+\rho+\frac{\rho^2}{(1-\rho)})=\frac{b}{1-\rho}.\end{aligned}$$

Hence, both cases coincide with the PS queue.

4.2 Priority queue

We now consider the case when the weight of the tagged customer grows large, i.e., $g_k \to \infty$. Hence, class k is prioritized in the limit. Then, the approximation simplifies to

$$\lim_{g_k\to\infty}\Bigg(b(1+\rho)$$
$$+\sum_{u_t=1}^{K}\lambda_{u_t}\mathbb{E}\Bigg[\frac{1}{2}\Bigg(\min\{B_{u_t},b\frac{g_{u_t}}{g_k}\}^2+\frac{g_k}{g_{u_t}}\min\{B_{u_t},b\frac{g_{u_t}}{g_k}\}^2\Bigg)-\big(b\frac{g_{u_t}}{g_k}\min\{B_{u_t},b\frac{g_{u_t}}{g_k}\}+B_{u_t}\min\{B_{u_t}\frac{g_k}{g_{u_t}},b\}\big)$$
$$+\frac{g_{u_t}}{g_k}bB_{u_t}\Bigg]+\frac{b(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{\mathbb{E}[B^2]}{g_k\sum_{\substack{j=1\\j\neq k}}^{K}\alpha_j\mathbb{E}[B_j^2]/g_j+\alpha_k\mathbb{E}[B_k^2]}\Bigg)$$
$$=b(1+\rho)+\mathbb{E}\Bigg[\sum_{\substack{u_t=1\\u_t\neq k}}^{K}\lambda_{u_t}\Bigg(\frac{1}{2}\big(0+0\big)-\big(0+bB_{u_t}\big)+0\Bigg)\Bigg]=b(1+\rho_k).$$

Note that the conditional sojourn time as $g_k \to \infty$ is known and its given by $b/(1-\rho_k)$. Since $1/(1-\rho_k) = \sum_{i=0}^{\infty}\rho_k^i$, we directly see that the approximation is the first order approximation of the exact expression. The relative error is equal to $100\%\left(b/(1-\rho_k)-b(1+\rho_k)\right)/b/(1-\rho_k)=\rho_k^2 100\%$, and we thus see that the relative error increases as the load of class k increases.

4.3 Monotonicity in the weights

It can be checked that the approximation for the mean conditional sojourn time of a tagged class-k customer, $\overline{S}_k^{INT}(\lambda,b)$, is decreasing in g_k and increasing in g_i, $i\neq k$.

This can be seen as follows. Conditioning on U_t we can write

$$\begin{aligned}\overline{S}_k^{INT}(\lambda,b) \;=\; & b(1+\rho)+\sum_{i=1,i\neq k}^{K}\lambda_i\mathbb{E}\Bigg[\frac{1}{2}\big(1+\frac{g_k}{g_i}\big)\min\{B_i,b\frac{g_i}{g_k}\}^2-\big(b\frac{g_i}{g_k}+\frac{g_k}{g_i}B_i\big)\min\{B_i,b\frac{g_i}{g_k}\}+b\frac{g_i}{g_k}B_i\Bigg]\\ &+\frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{b}{g_k}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K}\alpha_j\mathbb{E}[B_j^2]/g_j},\end{aligned}$$

where for $U_t=k$ we used that $\min\{B_k,b\}^2-\big(b+B_k\big)\min\{B_k,b\}+bB_k=0$.

Now, if $B_i\leq\frac{g_i}{g_k}b$, then

$$\frac{1}{2}\big(1+\frac{g_k}{g_i}\big)\min\{B_i,b\frac{g_i}{g_k}\}^2-\big(b\frac{g_i}{g_k}+\frac{g_k}{g_i}B_i\big)\min\{B_i,b\frac{g_i}{g_k}\}+b\frac{g_i}{g_k}B_i=\frac{1}{2}B_i^2(1-\frac{g_k}{g_i}),$$

which is decreasing in g_k and increasing in g_i. If $B_i>\frac{g_i}{g_k}b$, then

$$\frac{1}{2}\big(1+\frac{g_k}{g_i}\big)\min\{B_i,b\frac{g_i}{g_k}\}^2-\big(b\frac{g_i}{g_k}+\frac{g_k}{g_i}B_i\big)\min\{B_i,b\frac{g_i}{g_k}\}+b\frac{g_i}{g_k}B_i=\frac{1}{2}b^2\frac{g_i}{g_k}(1-\frac{g_i}{g_k})+bB_i(\frac{g_i}{g_k}-1),$$

which is decreasing in g_k and increasing in g_i (can be derived by taking the derivative and the fact that $B_i>\frac{g_i}{g_k}b$). The monotonicity of $\overline{S}_k^{INT}(\lambda,b)$ in g_k and g_i now follows immediately.

4.4 Uniformly bounded in the second moment

A very relevant property of processor sharing is that the mean sojourn time depends on the service time distribution only through its mean [Kelly 1979]. This has been an important argument to claim the interest of time-sharing disciplines with respect to more classical scheduling policies like FCFS. Indeed, the classical Pollaczek-Khinchine formula for the mean waiting time in a FCFS queue shows that it explodes as the second moment of the service time distribution grows large. For a DPS queue, Equation (1) does not allow to reach any conclusion regarding the dependence of the mean conditional sojourn time on the moments of the service time distribution.

It then becomes interesting to observe that the approximation (21) is uniformly bounded in the second moments of the service time distribution. To see this, we first note that $\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 \leq B_{U_t} b\frac{g_{U_t}}{g_k}$, which directly implies that the first three terms in (21) are uniformly bounded by a function that depends on the service requirements only through its first moment. We are now left with the heavy-traffic term $\dfrac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j \mathbb{E}[B_j^2]/g_j}$. Let j^* be such that $\mathbb{E}[B_{j^*}^2] \geq \mathbb{E}[B_j^2]$, $\forall j$. We then have

$$\frac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j \mathbb{E}[B_j^2]/g_j} = \frac{\sum_j \alpha_j \mathbb{E}[B_j^2]}{\sum_{j=1}^K \alpha_j \mathbb{E}[B_j^2]/g_j} \leq \frac{\mathbb{E}[B_{j^*}^2]}{\alpha_{j^*}\mathbb{E}[B_{j^*}^2]/g_{j^*}} = \frac{g_{j^*}}{\alpha_{j^*}}.$$

We thus finally conclude that (21) can be upper bounded by an expression that depends only on the first moment of the service time distributions. This indicates that the DPS queue provides a satisfactory performance even in the presence of service time distributions with a high variability.

4.5 Mean unconditional sojourn time

As a corollary of Proposition 4.2, we obtain the mean unconditional sojourn time of the tagged class-k customer.

COROLLARY 4.3. *The light and heavy-traffic interpolation (of order 2) of the mean unconditional sojourn time for a tagged class-k customer is given by*

$$\begin{aligned}
\overline{S}_k^{INT}(\lambda) &:= \int_0^\infty \overline{S}_k^{INT}(\lambda, b)\mathrm{d}F_k(b) \\
&= \mathbb{E}[B_k](1+\rho) + \lambda\mathbb{E}\left[\frac{1}{2}\big(1+\frac{g_k}{g_{U_t}}\big)\min\{B_{U_t}, B_k\frac{g_{U_t}}{g_k}\}^2 - \big(B_k\frac{g_{U_t}}{g_k} + \frac{g_k}{g_{U_t}}B_{U_t}\big)\min\{B_{U_t}, B_k\frac{g_{U_t}}{g_k}\} + B_k\frac{g_{U_t}}{g_k}B_{U_t}\right] \\
&\quad + \frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{\mathbb{E}[B_k]}{g_k}\frac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j\mathbb{E}[B_j^2]/g_j}. \qquad (22)
\end{aligned}$$

5. EXPONENTIAL SERVICE REQUIREMENTS

In this section we focus on the case in which the service requirements of the customers are exponentially distributed. We recall that a random variable B_i is exponentially distributed if $\mathbb{P}(B_i \leq b_i) = 1 - \mathrm{e}^{-b_i/\mathbb{E}[B_i]}$. In Section 5.1, we further simplify the expression for the light and heavy-traffic interpolation of the mean conditional and unconditional sojourn time and compare the latter for two classes of customers with the exact formulas as stated in Equations (5) and (6). In Section 5.2, we calculate the relative error (for different service requirements) and we verify that our approximation for the mean unconditional sojourn time for *an arbitrary customer* is exact when the service requirements of all customers are the same.

5.1 Mean conditional and unconditional sojourn time

In the case of exponentially distributed service requirements, our approximations for the mean conditional and unconditional sojourn time can be significantly simplified. This is a direct consequence of Proposition 4.2 and Corollary 4.3, respectively.

COROLLARY 5.1. *Assume class-k customers have an exponentially distributed service requirement with mean $1/\mu_k, k = 1, \ldots, K$. The light and heavy-traffic interpolation (of order 2) of the mean conditional sojourn time for a tagged class-k customer with service requirement b is given by*

$$\overline{S}_k^{INT}(\lambda, b) = b + \lambda\mathbb{E}[B]b + \lambda\sum_{j=1}^{K}\frac{\alpha_j}{\mu_j^2}\left(1-\frac{g_k}{g_j}\right)\left(1-\mathrm{e}^{-b\frac{g_j}{g_k}\mu_j}\right) + \frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{b}{g_k}\frac{\sum_{j=1}^{K}\alpha_j/\mu_j^2}{\sum_{j=1}^{K}\alpha_j/(\mu_j^2 g_j)}, \tag{23}$$

and the mean unconditional sojourn time is given by

$$\begin{aligned}\overline{S}_k^{INT}(\lambda) &:= \int_0^\infty \overline{S}_k^{INT}(\lambda, b)\mathrm{d}F_k(b)\\ &= \frac{1}{\mu_k} + \frac{1}{\mu_k}\lambda\mathbb{E}[B] + \lambda\sum_{j=1}^{K}\frac{\alpha_j}{\mu_j}\frac{(g_j-g_k)}{g_j\mu_j+g_k\mu_k} + \frac{(\lambda\mathbb{E}[B])^2}{(1-\lambda\mathbb{E}[B])}\frac{1}{g_k\mu_k}\frac{\sum_{j=1}^{K}\alpha_j/\mu_j^2}{\sum_{j=1}^{K}\alpha_j/(\mu_j^2 g_j)},\end{aligned} \tag{24}$$

where $\mathbb{E}[B] = \sum_{j=1}^{K}\alpha_j/\mu_j$.

See Appendix D for the proof.

In the case of two classes of customers, Fayolle et al [Fayolle et al. 1980] have closed-form expressions for the mean unconditional sojourn time, see Equation (5) and (6). Rewriting (24) for $K = 2$ in such a way so that the similarity with Equations (5) and (6) is clear, we obtain

$$\overline{S}_k^{INT}(\lambda) \;=\; \frac{1}{\mu_k(1-\rho)}\left(1+\rho^2\left(-1+\frac{\sum_{j=1}^{K}\alpha_j/\mu_j^2}{g_k\sum_{j=1}^{K}\alpha_j/(\mu_j^2 g_j)}\right)+\frac{\mu_k\rho_{-k}(g_{-k}-g_k)}{D}\frac{(1-\rho)D}{\mu_1 g_1+\mu_2 g_2}\right), \tag{25}$$

with $k = 1, 2$, $-k = \mathrm{mod}(k, 2) + 1$ and where $D = \mu_1 g_1(1-\rho_1) + \mu_2 g_2(1-\rho_2)$.

We directly observe that the difference with respect to the exact expression for the mean unconditional sojourn time ((5) and (6)) is in the terms

$$\rho^2\left(-1+\frac{\sum_{j=1}^{K}\alpha_j/\mu_j^2}{g_k\sum_{j=1}^{K}\alpha_j/(\mu_j^2 g_j)}\right) \quad \text{and} \quad \frac{(1-\rho)D}{\mu_1 g_1+\mu_2 g_2}.$$

It can easily be seen that our approximation is exact for the two extreme values of the traffic intensity, $\rho = 0$ and $\rho = 1$. That is, the expressions

$$\overline{S}_k^{INT}(0) = 1/\mu_k \quad \text{and} \quad \lim_{\lambda\to 1/\mathbb{E}[B]}(1-\rho)\overline{S}_k^{INT}(\lambda) = \frac{1}{\mu_k}\left(1+\frac{\mu_k\rho_{-k}(g_{-k}-g_k)}{D}\right)$$

are satisfied.

Let us denote by Rel.Error_k the relative error of a class-k customer, that is,

$$\text{Rel.Error}_k = \frac{\overline{S}_k(\lambda) - \overline{S}_k^{INT}(\lambda)}{\overline{S}_k(\lambda)}, \quad k = 1,\ 2.$$

Now, let us consider $g_1 + g_2 = 1$. We then obtain $\lim_{g_1 \uparrow 1} \text{Rel.Error}_1 = \rho_1^2 \cdot 100\%$ and

$$\lim_{g_1 \uparrow 1} \text{Rel.Error}_2 = \frac{\mu_2\rho_1 - (1-\rho_1)\left(\rho^2 \frac{\rho_1\mu_2}{\rho_2} + \mu_2\rho_1(1-\rho)\right)}{\mu_1(1-\rho_1) + \mu_2\rho_1} \cdot \ 100\%.$$

Hence, the relative error of class-1 customers (when $g_1 \uparrow 1$) increases as the load of class 1 increases but does not depend on the parameter of class 2. The same result was obtained in Section 4.2 for the mean conditional sojourn time for an arbitrary number of classes and general service requirements. Moreover, the absolute relative error of class-2 customers (when $g_1 \uparrow 1$) increases as the load of class 2 decreases.

In Figure 1 we plot the relative error of the mean unconditional sojourn time for $K = 2$ with respect to g_1. The parameters considered are $\rho_1 = 0.2, \rho_2 = 0.4, \mu_1 = 1, \mu_2 = 1, g_2 = 1 - g_1$ and from the formulas presented above we obtain $\lim_{g_1 \uparrow 1} \text{Rel. Error}_1 = 4\%$, $\lim_{g_1 \uparrow 1} \text{Rel. Error}_2 = -0.8\%$, $\lim_{g_1 \downarrow 0} \text{Rel. Error}_1 = -12.8\%$, $\lim_{g_1 \downarrow 0} \text{Rel. Error}_2 = 16\%$, which coincide with the extreme points in the figure.

5.2 Mean unconditional sojourn time for an arbitrary customer

In this section we discuss the relative error of the mean unconditional sojourn time for an *arbitrary customer*. We first calculate the error in case $K = 2$ and when μ_1 or μ_2 take extreme values. We then show that an approximation for the mean unconditional sojourn time of an arbitrary customer is exact when the service requirements of all customers are the same.

We denote by Rel.Error the relative error of an arbitrary customer, that is,

$$\text{Rel.Error} = \left(1 - \frac{\sum_{k=1}^{2} \alpha_k \overline{S}_k^{INT}(\lambda)}{\sum_{k=1}^{2} \alpha_k \overline{S}_k(\lambda)}\right) \cdot 100\%,$$

where $\overline{S}_1(\lambda)$ and $\overline{S}_2(\lambda)$ are given in Equations (5) and (6), respectively. Then, if we keep constant ρ_1 and ρ_2 we obtain

$$\lim_{\mu_2 \downarrow 0} \text{Rel.Error} = \frac{\rho_1 \left(\frac{\rho_2(g_2 - g_1)}{g_1}\left(\frac{1}{1-\rho_1} - 1 + \rho\right) - \rho^2\left(-1 + \frac{g_2}{g_1}\right)\right)}{\rho_1\left(1 + \frac{\rho_2(g_2-g_1)}{g_1(1-\rho_1)}\right) + \rho_2} \cdot 100\% \tag{26}$$

and $\lim_{\mu_1 \uparrow \infty} \text{Rel.Error} = \lim_{\mu_2 \downarrow 0} \text{Rel.Error}$. The intuition behind the latter equation can be seen as follows: having $\mu_1 \to \infty$ (and hence $\lambda_1 \to \infty$), i.e., having many class-1 arrivals of small size, is equivalent to having $\mu_2 \to 0$ (and hence $\lambda_2 \to 0$), i.e., having very few class-2 arrivals of large size.

In Figure 2 we plot the relative error of the mean unconditional sojourn time for *an arbitrary customer*. We fix ρ_1, ρ_2 and μ_1 and we let μ_2 and $\lambda_2 = \rho_2\mu_2$ change. The chosen parameters are $\rho_1 = 0.2, \rho_2 = 0.4, \mu_1 = 1, g_1 = 0.2, g_2 = 1 - g_1$. We observe that the results obtained from Equation (26), $\lim_{\mu_2 \downarrow 0} \text{Rel.Error} = -1.3\%$ and $\lim_{\mu_2 \uparrow \infty} \text{Rel.Error} = 6.4\%$, coincide with the extreme points in the figures.

We now show that our light-traffic approximation for the mean unconditional sojourn time of *an arbitrary customer* is exact under the assumption that the mean service requirements of all customers are the same, i.e., $\mathbb{E}[B_j] = 1/\mu, \forall j = 1, \ldots, K$. This result holds for an arbitrary number of classes.

As stated earlier, the mean unconditional sojourn time of *an arbitrary customer* is defined as $\overline{S}(\lambda) := \sum_{k=1}^{K} \alpha_k \overline{S}_k(\lambda)$. Since we assume exponentially distributed service requirements and $\mathbb{E}[B_k] = 1/\mu, \forall k = 1, \ldots, K$, the total number of customers in the system is distributed as that in a processor sharing

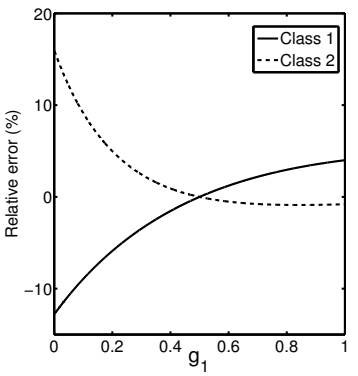

Fig. 1. Relative error for the mean unconditional sojourn time for $K = 2$ with respect to g_1.

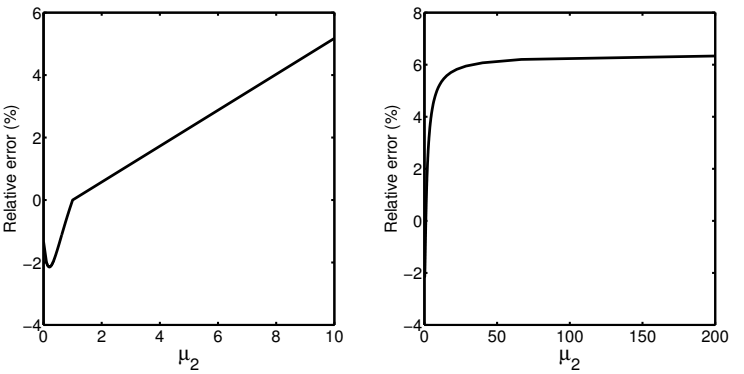

Fig. 2. Mean unconditional sojourn time for an arbitrary customer.

queue with arrival rate $\lambda = \sum_{k=1}^{K} \lambda_k$ and service rate μ. By Little's law, we therefore have that the total mean unconditional sojourn time is given by that of an $M/M/1$ queue, i.e, $\dfrac{1/\mu}{1-\rho}$.

For our light and heavy-traffic interpolation we have

$$\begin{aligned}\overline{S}^{INT}(\lambda) &= \sum_{k=1}^{K} \alpha_k \overline{S}_k^{INT}(\lambda) = \frac{1+\rho}{\mu} + \lambda \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{K} \frac{\alpha_j}{\mu} \frac{(g_j - g_k)}{g_j\mu + g_k\mu} + \frac{(\lambda/\mu)^2}{(1-\lambda/\mu)} \sum_{k=1}^{K} \alpha_k \frac{1}{\mu g_k} \frac{\sum_{j=1}^{K} \alpha_j/\mu^2}{\sum_{j=1}^{K} \alpha_j/(\mu^2 g_j)} \\ &= \frac{1+\rho}{\mu} + \frac{1}{\mu} \frac{\rho^2}{1-\rho} \sum_{k=1}^{K} \frac{\alpha_k}{g_k} \frac{1}{\sum_{j=1}^{K} \alpha_j/g_j} = \frac{1/\mu}{1-\rho}, \end{aligned} \tag{27}$$

where we used that

$$\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{K} \alpha_j \frac{(g_j - g_k)}{g_j\mu + g_k\mu} = \frac{1}{\mu} \sum_{k=1}^{K} \sum_{j=1}^{K} \alpha_k \alpha_j \frac{g_j - g_k}{g_j + g_k} = \frac{1}{\mu} \sum_{k=1}^{K} \sum_{j=1}^{K-1} \alpha_k \alpha_j \left(\frac{g_j - g_k}{g_j + g_k} + \frac{g_k - g_j}{g_j + g_k} \right) = 0.$$

Hence, the obtained light and heavy-traffic interpolation is exact when $\mathbb{E}[B_k] = 1/\mu, \forall k = 1, \ldots, K$.

In Figure 2 (*left*) we indeed observe that when $\mu_2 = 1$, so when the service requirements of both classes coincide, the relative error is 0, as proven in Equation (27).

6. NUMERICAL RESULTS

In this section we numerically investigate the accuracy of the approximations obtained in this paper. In Section 6.1 we consider the mean conditional sojourn time and in Section 6.2 the mean unconditional sojourn time, whose approximations are stated in Proposition 4.2 and Corollary 4.3, respectively.

As stated in Section 2, in [Fayolle et al. 1980] the authors obtain analytical expressions of the mean conditional and unconditional sojourn time under the assumption of exponentially distributed service requirements. For exponentially distributed service requirements, we will evaluate the accuracy of the approximations by comparing the exact formulas as obtained in [Fayolle et al. 1980], see Equations (2) and (4), with the approximations as given in (21) and (22). The expectations in (21) and (22) are calculated numerically using MATLAB's `integral2` command, which transforms the region of integration to a rectangular shape and subdivides it into smaller rectangular regions as needed.

In order to obtain a more complete understanding on the accuracy of the approximation, we will also consider hyperexponential and Pareto distributions. Hyperexponential and Pareto distributions have a decreasing hazard-rate, and their second moment can be made arbitrarily large. Because of these features they have been proposed as appropriate distributions to model service time distributions in the Internet.

We say that B_i has a hyperexponential distribution with m_i phases if

$$\mathbb{P}(B_i \leq b_i) = 1 - \sum_{k=1}^{m_i} p_{ik} \mathrm{e}^{(-b_i/\mathbb{E}[B_{ik}])}, \tag{28}$$

where p_{ik} is the probability that a class-i customer is exponentially distributed with mean $\mathbb{E}[B_{ik}]$. In particular, in Scenario 5 we consider a degenerate hyper-exponential distribution, for which we can easily obtain any value for the coefficient of variation without modifying the mean service time. In order to derive exact expressions for the mean sojourn time when the service requirements are hyperexponentially distributed, we make the observation that if classes $k = 1, \ldots, m_i$ are exponentially distributed (where class k has arrival rate λ_k and mean service requirement $\mathbb{E}[B_k]$) and have the same DPS weight, $g_1 = \ldots = g_{m_i}$, then they can be seen as a single (merged) class i with a hyperexponential distribution with parameters $p_{ik} = \lambda_k / \sum_{l=1}^{m_i} \lambda_l$ and $\mathbb{E}[B_{ik}] = \mathbb{E}[B_k]$, for each phase $k = 1, \ldots, m_i$. This allows us to calculate the performance with hyperexponential distribution using Equations (2) and (4) which are derived for exponential distributions.

We say that B_i has Pareto distribution with scale parameter c_i and shape parameter γ_i if

$$\mathbb{P}(B_i \leq b_i) = 1 - \left(\frac{1}{(1 + c_i b_i)} \right)^{\gamma_i}.$$

In order to derive exact expressions for the mean unconditional sojourn time we solved numerically Equation (1). However, the output was not stable enough and therefore we opted to simulate the DPS queue instead (using MATLAB). The simulation results are based on averaging 10 runs with each run comprising $5 \cdot 10^5$ busy periods. A busy period is defined as the interval of time between two consecutive time epochs when the system becomes empty, such points being regenerative points for the stochastic process of interest.

Note that the hyperexponential distribution satisfies the sufficient condition to hold admissibility (Equation (9)), whereas Pareto does not satisfy it; moments of an order higher than γ_i are unbounded. In the numerics we observe that even though condition (9) is not satisfied, the light-traffic interpolation approximation remains accurate.

Throughout this section the performance criteria will be the relative error. For instance, for the mean conditional sojourn time, we will calculate $100\% \times \frac{\overline{S}_k(\lambda,b) - \overline{S}_k^{INT}(\lambda,b)}{\overline{S}_k(\lambda,b)}$, and for the mean unconditional sojourn time $100\% \times \frac{\overline{S}_k(\lambda) - \overline{S}_k^{INT}(\lambda)}{\overline{S}_k(\lambda)}$.

Before explaining in detail the numerical results we have obtained, we summarize our main conclusions:

—The approximation is accurate over a broad range of parameter values.

—For a given set of parameters, the relative error for the mean conditional sojourn time increases as the service requirement of the tagged customer increases.

—The error increases as the disparity among the weights increases.

—For any given scenario, the largest relative error occurs in an intermediate load between 0 and 1.

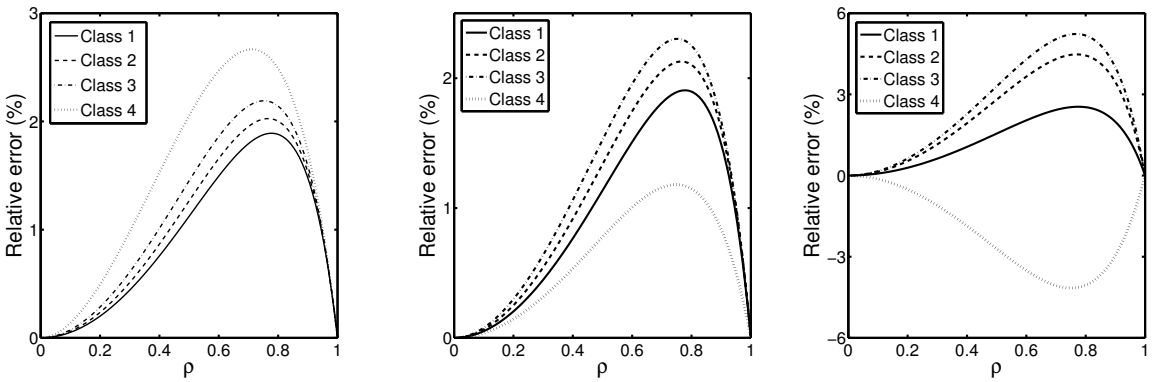

Fig. 3. Scenario 1: Relative error for the mean conditional sojourn time for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

—The largest relative errors for the mean conditional sojourn time occur for service requirements b that are very unlikely to occur. This also explains the high accuracy of our approximation for the mean unconditional sojourn time.

—Although the Pareto distribution does not satisfy the admissibility condition (Equation (9)), the light-traffic interpolation approximation remains accurate.

—We compare our approximation to that obtained in [van Kessel et al. 2005] and conclude that our approximation outperforms that of [van Kessel et al. 2005].

—We observe that our approximation works well across different values of the coefficient of variation.

6.1 Conditional sojourn time

In this section we measure the accuracy of the mean conditional sojourn time given in Proposition 4.2.

Scenario 1. In Figure 3 we consider four classes $K = 4$ with exponentially distributed service requirements. The parameters of the classes are fixed, and we vary the total arrival rate in order for the load to cover the range of stable values. We consider $\mathbb{E}[B_1] = 2$, $\mathbb{E}[B_2] = 5$, $\mathbb{E}[B_3] = 7$, $\mathbb{E}[B_4] = 10$, $g_1 = 30$, $g_2 = 25$, $g_3 = 20$, $g_4 = 10$, and $\alpha_1 = 10/36$, $\alpha_2 = 5/36$, $\alpha_3 = 8/36$, $\alpha_4 = 13/36$ such that $\lambda_i = \alpha_i * \lambda$, $i = 1, \ldots, 4$, where λ is the total arrival rate. In Figure 3 we plot the relative error of our approximation for the mean conditional sojourn time of a tagged class-i customer, for $i = 1, \ldots, 4$, where the size of the tagged class-i customer, b_i, is selected such that the probability of the event is $\mathbb{P}(B_i \leq b_i) = 0.01$, $\mathbb{P}(B_i \leq b_i) = 0.50$ and $\mathbb{P}(B_i \leq b_i) = 0.99$, respectively. As can be seen, the relative error for the mean conditional sojourn time remains small and always below 6%.

Scenario 2. In Figure 4 we consider two classes $K = 2$ with exponentially distributed service requirements. We fixed the parameters $\mathbb{E}[B_1] = 2$, $\mathbb{E}[B_2] = 1$, $g_1 = 1$, $g_2 = 3$, $\alpha_1 = 0.415$, $\alpha_2 = 0.585$ and $\lambda_i = \alpha_i * \lambda$. We let the service requirement of the class-i tagged customer span between 0 and $b_{i,max}$ where $\mathbb{P}(B_i \leq b_{\text{i,max}}) = 0.99$ and for each b we plot the largest absolute relative error that can be found for a $\rho \in [0, 1)$. We observe a largest error of at most 6%.

Scenario 3. In Figure 5 we consider again two classes with exponentially distributed service requirements. As parameters we fix: $\mathbb{E}[B_1] = 2$, $\mathbb{E}[B_2] = 1$, $\lambda_1 = 0.2$, $\lambda_2 = 1.5\lambda_1$ and $b = 1$. We chose $g_2 = 1 - g_1$

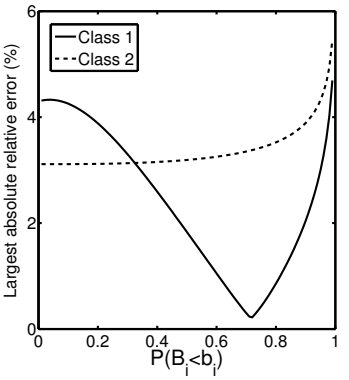

Fig. 4. Scenario 2: Largest absolute relative error for the mean conditional sojourn time as a function of $\mathbb{P}(B_i \leq b_i)$, $i = 1, 2$.

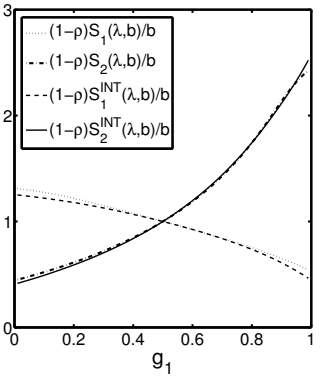

Fig. 5. Scenario 3: Mean conditional sojourn time as a function of g_1.

and let g_1 vary on the horizontal axis. In the figure we plot the mean conditional sojourn time and our approximation. We see that the property stated in Section 4.3 is satisfied, namely as g_1 increases $\overline{S}_1^{INT}(\lambda, b)$ decreases and $\overline{S}_2^{INT}(\lambda, b)$ increases. Besides, it can be observed from the figure that the approximation looses accuracy as one class is given more priority, i.e., $g_1 \to 0$ or $g_1 \to 1$.

Scenario 4. In Figure 6 we consider two classes with hyperexponential distributed service requirements with $\mathbb{E}[B_1] = 11/3$, $\mathbb{E}[B_2] = 44/3$. Each of the hyperexponential distributions has 3 phases. The parameters are as follows: for class 1 we take $\mathbb{E}[B_{11}] = 3.5$, $\mathbb{E}[B_{12}] = 2$, $\mathbb{E}[B_{13}] = 5$, $p_{11} = 10/21$, $p_{12} = 5/21$, $p_{13} = 6/21$, and for class 2 we take $\mathbb{E}[B_{21}] = 10$, $\mathbb{E}[B_{22}] = 15$, $\mathbb{E}[B_{23}] = 20$, $p_{21} = 4/15$, $p_{22} = 8/15$, $p_{23} = 3/15$. The weights are set to $g_1 = 2$ and $g_2 = 5$. We assume that an arriving customer is of class 1 (class 2) with probability $\alpha_1 = 7/12$ ($\alpha_2 = 5/12$). As in *Scenario 1*, we select the service requirement of the tagged customer such that $\mathbb{P}(B_i \leq b_i) = 0.01, 0.5$ and 0.99. We see that the error increases as the size of the tagged customer increases. However it is remarkable how accurate the approximation is.

In Figure 7 we consider *Scenario 4* with weights $g_1 = 2, g_2 = 5$ (left) and $g_1 = 5, g_2 = 2$ (right), respectively. We vary the service requirement of the class-i tagged customer between 0 and $b_{i,max}$ where $\mathbb{P}(B_i \leq b_{\text{i,max}}) = 0.99$ and for each b we plot the largest absolute relative error that can be found for a $\rho \in [0, 1)$. It can be observed that the largest absolute relative error is smaller as more priority is given to the class with small mean service requirements. In Figure 7 (left) the largest absolute relative error is of the order of 9% for the class with the smallest weight and of the order of 2% for the class with the highest weight. While in Figure 7 (right) the largest absolute relative error is of the order of 1.7% for the class with the smallest weight and of the order of 2.5% for the class with the highest weight.

6.2 Unconditional sojourn time

In this section we evaluate the accuracy of the mean unconditional sojourn time given in Corollary 4.3.

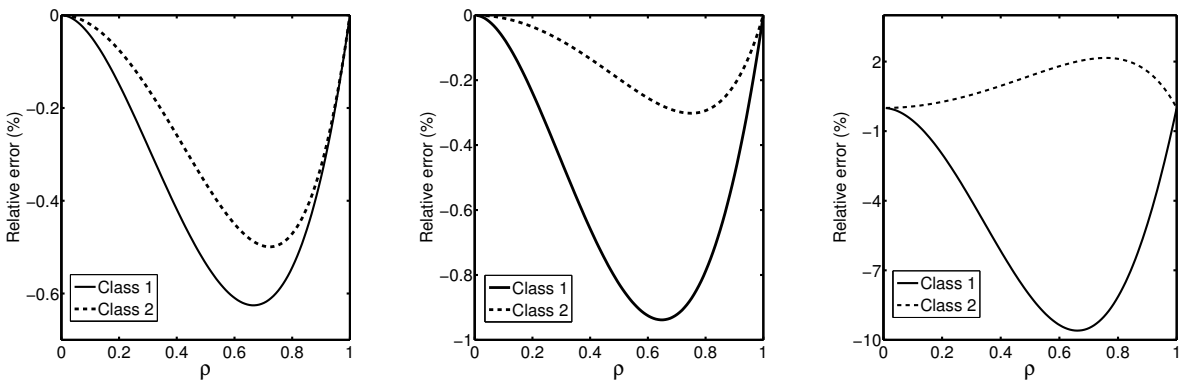

Fig. 6. Scenario 4: Relative error for the mean conditional sojourn time for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

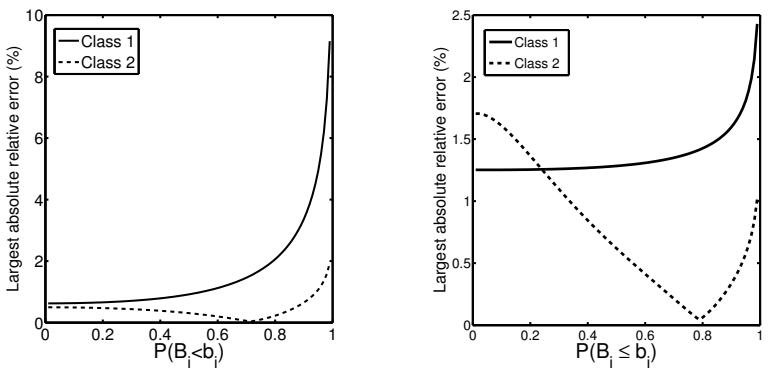

Fig. 7. Scenario 4: Largest absolute relative error for the mean conditional sojourn time as a function of $\mathbb{P}(B_i \leq b_i)$, $i = 1, 2$ with weights $g_1 = 2, g_2 = 5$ (left) and $g_1 = 5, g_2 = 2$ (right).

In Figure 8 we consider the same parameter setting as in *Scenario 1*, and we observe that the largest relative error for the mean unconditional sojourn time is less than 3.5%.

In Figure 9 we consider two classes with hyper-exponentially distributed service requirements. The parameters are the ones considered in *Scenario 4*. We plot the relative error of the mean unconditional sojourn time for our approximation. We conclude that our approximation works very well. The largest relative error for the mean unconditional sojourn time is around 3%. We compare our approximation to that obtained in [van Kessel et al. 2005]. In the latter, the unconditional sojourn time was approximated by expressions obtained when one of the classes lives on a relatively faster time scale than the other class. Under this scenario, class 1 represents the class of a vast majority of customers with small service requirements while class 2 represents the tiny fraction of customers with huge service requirements. Therefore, the approximation as given in [van Kessel et al. 2005] is $\mathbb{E}[S_1(\lambda)] \approx \left(\frac{g_2}{g_1}\frac{\rho_2}{1-\rho} + 1\right)\frac{\mathbb{E}[B_1]}{1-\rho_1}$ and $\mathbb{E}[S_2(\lambda)] \approx \frac{\rho_2}{1-\rho}$. In Figure 9 we plot the mean unconditional

sojourn time of [van Kessel et al. 2005]. We observe that the largest relative error is 17%. We conclude that our approximation outperforms that of [van Kessel et al. 2005].

As pointed out in the beginning of the section, we observe that the relative error for the mean unconditional sojourn time tends to be smaller than the ones observed for the mean conditional sojourn time. This can be explained by noting that the largest errors in the mean conditional sojourn time tend to occur for service requirements that happen with a very low probability.

In Figure 10 we consider two classes with hyper-exponentially distributed service requirements. The parameters are the same as in *Scenario 4*. We chose $g_2 = 1 - g_1$ and let g_1 vary on the horizontal axis. For each given g_1 we calculate the largest absolute relative error for the mean unconditional sojourn time as we let ρ range from 0 to 1. We observe that the relative error for the unconditional sojourn time is at most of 30%, and that this happens when class 2 receives full priority.

Scenario 5. In Figure 11 we consider 2 classes of customers. Class-1 customers' service requirements follow an exponential distribution of rate μ_1, while class-2 customers' service requirements follow a hyper-exponential distribution as defined in Equation (28) with parameters $m_2 = 2, p_{21} = p, p_{22} = 1 - p, \mathbb{E}[B_{21}] = 1/(\mu_2 p)$ and $\mathbb{E}[B_{22}] = 0$. The latter distribution is referred to as a degenerate hyper-exponential distribution with $\mathbb{E}[B_2] = 1/\mu_2$ and $\mathbb{E}[B_2^2] = \frac{2p}{(p\mu_2)^2} = \frac{2}{p\mu_2^2}$. We then easily obtain that the coefficient of variation satisfies $C_{B_2}^2 = \dfrac{2}{p} - 1$, and conclude that $C_{B_2} \in [1, \infty)$ as $p \in [0, 1]$.

We consider $\rho = 0.7$, $g_1 = 1$, $g_2 = 5$, $\alpha_1 = 1/4$, $\alpha_2 = 3/4$, $\mu_1 = 1$ and $\mu_2 = 1$.

In Figure 11 we plot the relative error of the mean unconditional sojourn time for an *arbitrary* customer with respect to p for our approximation. We observe that our approximation works well across all values of p. In the limit $p \downarrow 0$ we obtain the following analytical expression for the relative error

$$\lim_{p \downarrow 0} \text{Rel.Error} = \left(1 - \frac{1 + \alpha_1 \left(\dfrac{g_2}{g_1} - 1 \right) \left(\rho^2 + \rho_2 (1 - \rho) \right)}{1 + \alpha_1 \left(\dfrac{g_2}{g_1} - 1 \right) \dfrac{\rho_2}{1 - \rho_1}} \right) \cdot 100\%. \tag{29}$$

For the parameters of Figure 11, the latter is equal to -0.6806% (see also the curve in Figure 11). This implies that for very high coefficient of variation $(p \downarrow 0)$ the performance of our approximation is very good. From Equation (29) we note that even though the coefficient of variation explodes as $p \downarrow 0$, the relative error of our approximations remains bounded. In the limit $p \uparrow 1$, our approximation is exact since in this particular case class-2 customers follow an exponential distribution of rate μ_2 and under the assumption $\mu_1 = \mu_2$ we proved in Equation (27) that our approximation becomes exact. In Figure 11 we also plot the approximation as obtained in [van Kessel et al. 2005]. We note that as $p \downarrow 0$, class-2 customers arrive very rarely and are huge. Hence, the approximation of [van Kessel et al. 2005] becomes exact as $p \downarrow 0$. For $p > 0$ the absolute relative error is monotone increasing taking the value 60% at $p \uparrow 1$.

In the following two scenarios we consider Pareto distributed serve requirements. As mentioned in the beginning of the section Pareto does not satisfy the admissibility condition (Equation (9)). However, we will observe that the light-traffic interpolation remains accurate.

Scenario 6. In Figure 12 we consider Pareto distributed service requirements. We consider four classes with $c_1 = 1/4, c_2 = 1/10, c_3 = 1/14, c_4 = 1/20$ and $\gamma_1 = 3, \gamma_2 = 3, \gamma_3 = 3, \gamma_4 = 3$, such that,

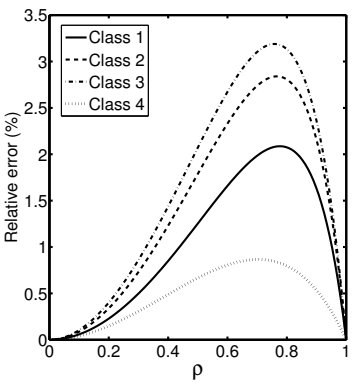

Fig. 8. Scenario 1: Relative error for the mean unconditional sojourn time.

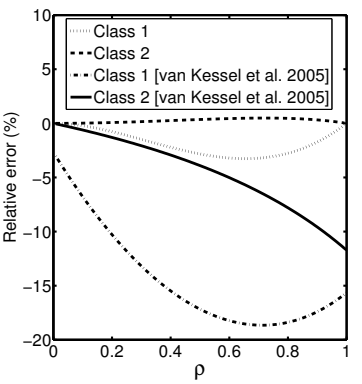

Fig. 9. Scenario 4: Relative error for the mean unconditional sojourn time.

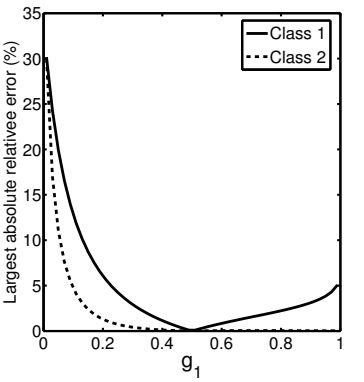

Fig. 10. Scenario 4: Largest absolute relative error for the mean unconditional sojourn time with respect to the weight g_1.

$\mathbb{E}[B_1] = 2$, $\mathbb{E}[B_2] = 5$, $\mathbb{E}[B_3] = 7$, $\mathbb{E}[B_4] = 10$. The weights are set to $g_1 = 30$, $g_2 = 25$, $g_3 = 20$, $g_4 = 10$ and $\alpha_1 = 10/36$, $\alpha_2 = 5/36$, $\alpha_3 = 8/36$, $\alpha_4 = 13/36$ such that $\lambda_i = \alpha_i * \lambda$, $i = 1, \ldots, 4$, where λ is the total arrival rate. Notice that $\mathbb{E}[B_i], g_i, \alpha_i, i = 1, \ldots, 4$ are as in Scenario 1. We note that the point $\rho = 1$ is obtained from the heavy-traffic condition and is therefore exact. For $\rho \neq 1$, we simulated the system in order to compare our approximation. We conclude that the largest relative error for the mean unconditional sojourn time is around 5%.

Scenario 7. In Figure 13 we consider Pareto distributed service requirements. We consider two classes with $c_1 = 3/22, c_2 = 3/88$ and $\gamma_1 = 3, \gamma_2 = 3$, such that, $\mathbb{E}[B_1] = 11/3$, $\mathbb{E}[B_2] = 44/3$ and the weights are set to $g_1 = 2$ and $g_2 = 5$. We assume that an arriving customer is of class 1 (class 2) with probability $\alpha_1 = 7/12$ ($\alpha_2 = 5/12$). Notice that $\mathbb{E}[B_i], g_i, \alpha_i, i = 1, 2$ are as in Scenario 4. We note that the point $\rho = 1$ is obtained from the heavy-traffic condition and is therefore exact. For $\rho \neq 1$, we simulated the system in order to compare our approximation. We conclude that the largest relative error for the mean unconditional sojourn time is less than 5%.

Acknowledgements

The authors are thankful to Professor A.M. Makowski (University of Maryland at College Park) for helpful discussions.
Research partially supported by the French "Agence Nationale de la Recherche (ANR)" through the project ANR JCJC RACON.

REFERENCES

ALTMAN, E., AVRACHENKOV, K., AND AYESTA, U. 2006. A survey on discriminatory processor sharing. *Queueing systems 53,* 1-2, 53–63.

ALTMAN, E., JIMENEZ, T., AND KOFMAN, D. 2004. DPS queues with stationary ergodic service times and the performance of TCP in overload. In *Proceedings of IEEE INFOCOM*.

AVRACHENKOV, K., AYESTA, U., BROWN, P., AND NÚÑEZ-QUEIJA, R. 2005. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM*.

BORST, S., NÚÑEZ-QUEIJA, R., AND ZWART, A. 2006. Sojourn time asymptotics in processor sharing queues. *Queueing Systems 53,* 1–2, 31–51.

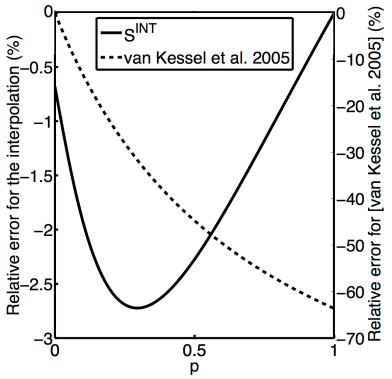

Fig. 11. Scenario 5: Relative error for the mean unconditional sojourn time of an *arbitrary* customer for our approximation and for the approximation as obtained in [van Kessel et al. 2005].

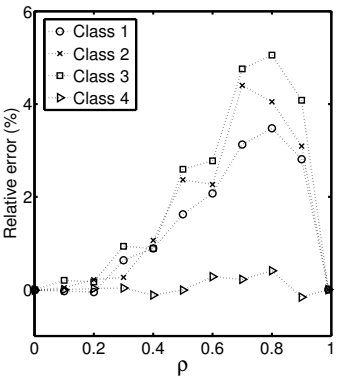

Fig. 12. Scenario 6: Relative error for the mean unconditional sojourn time.

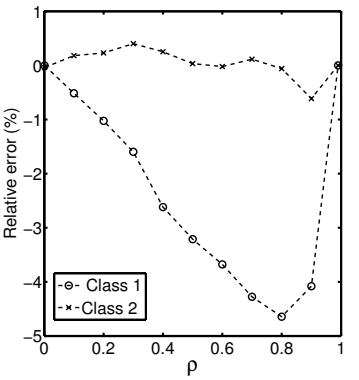

Fig. 13. Scenario 7: Relative error for the mean unconditional sojourn time.

BORST, S., VAN OOTEGHEM, D., AND ZWART, A. 2005. Tail asymptotics for discriminatory processor sharing queues with heavy-tailed service requirements. *Performance Evaluation 61,* 2–3, 281–298.

BOXMA, O., HEGDE, N., AND NÚÑEZ-QUEIJA, R. 2006. Exact and approximate analysis of sojourn times in finite discriminatory processor sharing queues. *AEU International Journal on Electronic Communications 60*, 109–115.

BU, T. AND TOWSLEY, D. 2001. Fixed point approximation for TCP behaviour in an AQM network. In *Proceedings of ACM SIGMETRICS/Performance*. 216–225.

CHEUNG, S., VAN DEN BERG, J., BOUCHERIE, R., LITJENS, R., AND ROIJERS, F. 2005. An analytical packet/flow-level modelling approach for wireless LANs with quality-of-service support. In *Proceedings of ITC-19*.

FAYOLLE, G., MITRANI, I., AND IASNOGORODSKI, R. 1980. Sharing a processor among many job classes. *Journal of the ACM 27,* 3, 519–532.

FREDJ, S. B., BONALD, T., PROUTIERE, A., REGNIE, G., AND ROBERTS, J. 2001. Statistical bandwidth sharing: A study of congestion at flow level. In *SIGCOMM*. 111–122.

GRISHECHKIN, S. 1992. On a relationship between processor sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Prob. 24,* 3, 653–698.

HASSIN, R. AND HAVIV, M. 2003. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston etc.

HAYEL, Y. AND TUFFIN, B. 2005. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *Proceedings of Networking*.

IZAGIRRE, A. 2015. Interpolation approximations for steady-state performance measures. Ph.D. thesis, INSA Toulouse.

IZAGIRRE, A., AYESTA, U., AND VERLOOP, I.M. 2014. Sojourn time approximations in a multi-class time-sharing system. *IEEE INFOCOM*.

IZAGIRRE, A., AYESTA, U., AND VERLOOP, I.M. 2015. Interpolation approximations for the steady-state distribution in multi-class resource-sharing systems. *Performance Evaluation http://dx.doi.org/10.1016/j.peva.2015.06.005*.

KELLY, F. 1979. *Stochastic Networks and Reversibility*. Wiley, Chichester.

KELLY, F. 1997. Charging and rate control for elastic traffic. *European Transactions on Telecommunications 8*, 33–37.

KHERANI, A. AND NÚÑEZ-QUEIJA, R. 2006. TCP as an implementation of age-based scheduling: fairness and performance. In *Proceedings of IEEE INFOCOM*.

KLEINROCK, L. 1967. Time-shared systems: A theoretical treatment. *Journal of the ACM 14,* 2, 242–261.

KLEINROCK, L. 1976. *Queueing Systems, vol. 2*. John Wiley and Sons.

REGE, K. AND SENGUPTA, B. 1996. Queue length distribution for the discriminatory processor sharing queue. *Operations Research 44,* 4, 653–657.

REIMAN, M. AND SIMON, B. 1988a. An interpolation approximation for queueing systems with Poisson input. *Operations Research 36*, 454–469.

REIMAN, M. AND SIMON, B. 1988b. Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models 4*, 191–233.

REIMAN, M. AND SIMON, B. 1989. Open queueing systems in light traffic. *Oper. Res. 14*, 26–59.

ROBERTS, J. 2004. A survey on statistical bandwidth sharing. *Computer Networks 45*, 319–332.

VAN KESSEL, G., NÚÑEZ-QUEIJA, R., AND BORST, S. 2005. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM*.

VERLOOP, I.M., AYESTA, U., AND NÚÑEZ-QUEIJA, R. 2011. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Operations Research 59,* 3, 648–660.

WALRAND, J. 1988. An introduction to queueing networks. In *Prentice-Hall, Englewood Cliffs, NJ*.

WU, Y., BUI, L., AND JOHARI, R. 2012. Heavy traffic approximation of equilibria in resource sharing games. *IEEE Journal on Selected Areas in Communications 30,* 11, 2200–2209.

YASHKOV, S. 1987. Processor sharing queues: Some progress in analysis. *Queueing Systems 2*, 1–17.

Appendix A: Light-traffic approach

We provide an intuitive approach of how to obtain the zeroth and first light-traffic derivatives. This approach is based on the analysis of J. Walrand in [Walrand 1988, Chapter 6.3.]. Higher order light-traffic derivatives can be obtained in a similar way.

Consider a system that starts at time $-Z$ and that keeps going until time T, being $Z, T > 0$ given. Let $S_k(\lambda, b, -Z)$ denote in this case the sojourn time of the tagged class-k customer who arrives in the system at time $t = 0$ and note that $\lim_{Z\to\infty} S_k(\lambda, b, -Z) = S_k(\lambda, b)$. Let $A(s, t)$ denote the number of arrivals in the interval $[s, t)$ in addition to the tagged customer who is assumed to arrive at time 0. Throughout this section we assume that the limits (with respect to Z and T) and expectations can be interchanged. We then have

$$\mathbb{E}\left[\min\{S_k(\lambda, b, -Z), T\}\right] = \sum_{a=0}^{\infty} \mathbb{E}\left[\min\{S_k(\lambda, b, -Z), T\}\Big| A(-Z, T) = a\right] \cdot \frac{(\lambda(T+Z))^a}{a!} \mathrm{e}^{-\lambda(T+Z)}, \tag{30}$$

where $\mathbb{E}\left[\min\{S_k(\lambda, b, -Z), T\}\Big| A(-Z, T) = a\right]$ is the expected minimum between the sojourn time and T, conditioned that there are exactly a arrivals. Evaluating it at $\lambda = 0$ gives

$$\mathbb{E}\left[\min\{S_k(\lambda, b, -Z), T\}\right]\Big|_{\lambda=0} = \mathbb{E}\left[\min\{S_k(0, b, -Z\}, T\}\Big| A(-Z, T) = 0\right], \tag{31}$$

and now taking the limit $Z, T \to \infty$ we obtain the zeroth light-traffic derivative

$$\overline{S}_k(0, b) := \lim_{Z,T\to\infty} \mathbb{E}\left[\min\{S_k(\lambda, b, -Z), T\}\right]\Big|_{\lambda=0} = \mathbb{E}\left[S_k(0, b)\Big| A(-\infty, \infty) = 0\right],$$

where the second equality follows from (31).

Next, consider the derivative with respect to λ in Equation (30) and evaluate it at $\lambda = 0$. This gives

$$\frac{\partial}{\partial\lambda}\mathbb{E}\left[\min\{S_k(\lambda,b,-Z),T\}\right]\Big|_{\lambda=0}$$
$$= -\mathbb{E}\left[\min\{S_k(0,b,-Z),T\}\Big|A(-Z,T)=0\right]\cdot(T+Z)+\mathbb{E}\left[\min\{S_k(0,b,-Z),T\}\Big|A(-Z,T)=1\right]\cdot(T+Z)$$
$$= \int_{-Z}^{T}\left(\mathbb{E}\left[\min\{S_k(0,b,-Z),T\}\Big|A(-Z,T)=1,\tau=t\right]-\mathbb{E}\left[\min\{S_k(0,b,-Z),T\}\Big|A(-Z,T)=0\right]\right)dt, \quad (32)$$

where τ is the arrival time of the first customer. The second equality holds because the arrivals follow a Poisson process. Hence given that the number of arrivals in $[-Z,T)$ is one ($A(-Z,T)=1$), we have that τ is uniformly distributed on $[-Z,T)$.

Now taking $Z,T\to\infty$ we obtain the first light-traffic derivative

$$\overline{S}_k^{(1)}(0,b) := \lim_{Z,T\to\infty}\frac{\partial}{\partial\lambda}\mathbb{E}\left[\min\{S_k(\lambda,b,-Z),T\}\right]\Big|_{\lambda=0}$$
$$= \int_{-\infty}^{\infty}\left(\mathbb{E}\left[S_k(0,b)\Big|A(-\infty,\infty)=1,\tau=t\right]-\mathbb{E}\left[S_k(0,b)\Big|A(-\infty,\infty)=0\right]\right)dt,$$

where the second equality follows from (32).

Appendix B: Proof of Lemma 3.1

To calculate $\overline{S}_k^{(1)}(0,b)$ we need to calculate $\int_{-\infty}^{\infty}\mathbb{E}[S_{k,t,U_t,B_{U_t}}]\mathrm{d}t$, where $S_{k,t,u_t,b_{u_t}}$ is as given in Equation (13). We first focus on the calculation corresponding to the first term of (13), that is, the case when $t\le 0\le t+B_{U_t}$ and $t<\frac{g_{U_t}}{g_k}b-B_{U_t}$, (where the inequalities of the random variables hold sample-path wise). We have

$$\int_{-\infty}^{0}\mathbb{E}\left[\mathbf{1}\left[-B_{U_t}\le t<\frac{g_{U_t}}{g_k}b-B_{U_t}\right](t+B_{U_t}+b)\right]\mathrm{d}t=\int_{0}^{\infty}\mathbb{E}\left[\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right](-t+B_{U_t}+b)\right]\mathrm{d}t$$
$$=\mathbb{E}\left[\int_0^{\infty}\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right](-t+B_{U_t}+b)\,\mathrm{d}t\right],$$

as we make use of Tonelli's Theorem. It follows that

$$\int_0^{\infty}\mathbf{1}\left[B_{U_t}\ge t>B_{U_t}-\frac{g_{U_t}}{g_k}b\right](-t+B_{U_t}+b)\,\mathrm{d}t$$
$$=\int_{\left(B_{U_t}-\frac{g_{U_t}}{g_k}b\right)^+}^{B_{U_t}}(-t+B_{U_t}+b)\,\mathrm{d}t=\left[-\frac{t^2}{2}+B_{U_t}t+bt\right]_{\left(B_{U_t}-\frac{g_{U_t}}{g_k}b\right)^+}^{B_{U_t}}. \quad (33)$$

We can now consider two cases. If $B_{U_t} - \frac{g_{U_t}}{g_k} b > 0$, then Equation (33) is equal to

$$\begin{aligned}
&\left[-\frac{t^2}{2} + B_{U_t} t + bt \right]_{B_{U_t} - \frac{g_{U_t}}{g_k} b}^{B_{U_t}} \\
&= -\frac{B_{U_t}^2}{2} + B_{U_t} B_{U_t} + b B_{U_t} - \left(-\frac{\left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right)^2}{2} + B_{U_t} \left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right) + b \left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right) \right) \\
&= \frac{1}{2} \left(\frac{g_{U_t} b}{g_k} \right)^2 + \frac{g_{U_t} b^2}{g_k}.
\end{aligned}$$

If $B_{U_t} - \frac{g_{U_t}}{g_k} b < 0$, then Equation (33) is equal to

$$\left[-\frac{t^2}{2} + B_{U_t} t + bt \right]_0^{B_{U_t}} = -\frac{B_{U_t}^2}{2} + B_{U_t} B_{U_t} + b B_{U_t} \ = \frac{B_{U_t}^2}{2} + b B_{U_t}.$$

We thus obtain

$$\mathbb{E}\left[\int_0^\infty \mathbf{1}\left[B_{U_t} \geq t > B_{U_t} - \frac{g_{U_t}}{g_k} b \right] (-t + B_{U_t} + b)\, \mathrm{d}t \right] = \mathbb{E}\left[\frac{1}{2} \min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 + b \min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} \right]. \quad (34)$$

Second, we focus on the calculation corresponding to the second term of (13), that is, the case when $t \leq 0 \leq t + B_{U_t}$ and $\frac{g_{U_t}}{g_k} b - t \leq B_{U_t}$. We have

$$\begin{aligned}
\int_{-\infty}^0 \mathbb{E}\left[\mathbf{1}\left[\frac{g_{U_t}}{g_k} b - B_{U_t} \leq t \right] \frac{g_k + g_{U_t}}{g_k} b \right] \mathrm{d}t \ &= \ \int_0^\infty \mathbb{E}\left[\mathbf{1}\left[\frac{g_{U_t}}{g_k} b - B_{U_t} \leq -t \right] \frac{g_k + g_{U_t}}{g_k} b \right] \mathrm{d}t \\
&= \ \mathbb{E}\left[\int_0^\infty \mathbf{1}\left[\frac{g_{U_t}}{g_k} b - B_{U_t} \leq -t \right] \frac{g_k + g_{U_t}}{g_k} b \mathrm{d}t \right],
\end{aligned}$$

as we make use of Tonelli's Theorem. It follows that

$$\int_0^\infty \mathbf{1}\left[t \leq B_{U_t} - \frac{g_{U_t}}{g_k} b \right] \frac{g_k + g_{U_t}}{g_k} b \mathrm{d}t = \frac{g_k + g_{U_t}}{g_k} \int_0^{\left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right)^+} b \mathrm{d}t = b \frac{g_k + g_{U_t}}{g_k} \left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right)^+. \quad (35)$$

We can now consider two cases. If $B_{U_t} - \frac{g_{U_t}}{g_k} b > 0$, then Equation (35) is equal to $b \frac{g_k + g_{U_t}}{g_k} \left(B_{U_t} - \frac{g_{U_t}}{g_k} b \right)$. If $B_{U_t} - \frac{g_{U_t}}{g_k} b \leq 0$, then Equation (35) is equal to 0. We thus obtain

$$\mathbb{E}\left[\int_0^\infty \mathbf{1}\left[\frac{g_{U_t}}{g_k} b - B_{U_t} \leq -t \right] \frac{g_k + g_{U_t}}{g_k} b \mathrm{d}t \right] = \mathbb{E}\left[\frac{g_k + g_{U_t}}{g_k} b \left(B_{U_t} - \min\{B_{U_t}, \frac{g_{U_t}}{g_k} b\} \right) \right]. \quad (36)$$

Third, we focus on the subtraction between the third term of (13), that is, the case when $t + B_{U_t} < 0$, and $\mathbb{E}\left[S_k(0, b) \middle| A = 0 \right] = b$. We then have

$$\int_{-\infty}^0 \mathbb{E}[\mathbf{1}\left[t < -B_{U_t} \right] b - b] \mathrm{d}t = \int_0^\infty \mathbb{E}[\mathbf{1}\left[-t < -B_{U_t} \right] b - b] \mathrm{d}t = \mathbb{E}\left[b \int_0^\infty (\mathbf{1}\left[B_{U_t} < t \right] - 1) \mathrm{d}t \right]$$

as we make use of Tonelli's Theorem. It follows that

$$b\int_0^\infty (\mathbf{1}\left[B_{U_t} < t\right] - 1)\mathrm{d}t = b\int_0^\infty -\mathbf{1}\left[B_{U_t} > t\right]\mathrm{d}t = -b\int_0^{B_{U_t}} \mathrm{d}t = -bB_{U_t}.$$

We thus obtain

$$\mathbb{E}\left[\int_0^\infty (\mathbf{1}\left[B_{U_t} < t\right] b - b)\mathrm{d}t\right] = -b\mathbb{E}[B_{U_t}]. \tag{37}$$

Fourth, we focus on the calculation corresponding to the fourth term of (13), that is, the case when $0 < t < b$ and $\frac{b-t}{g_k} > \frac{B_{U_t}}{g_{U_t}}$. We have

$$\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[t < b - \frac{g_k B_{U_t}}{g_{u_t}}\right](b + B_{U_t})\right]\mathrm{d}t = \mathbb{E}\left[\int_0^\infty \mathbf{1}\left[t < b - \frac{g_k B_{U_t}}{g_{U_t}}\right](b + B_{U_t})\,\mathrm{d}t\right],$$

as we make use of Tonelli's Theorem. It follows that

$$\int_0^\infty \mathbf{1}\left[t < b - \frac{g_k B_{U_t}}{g_{U_t}}\right](b + B_{U_t})\,\mathrm{d}t = \int_0^{\left(b - \frac{g_k B_{U_t}}{g_{U_t}}\right)^+} (b + B_{U_t})\,\mathrm{d}t. \tag{38}$$

If $b - \frac{g_k B_{U_t}}{g_{U_t}} > 0$ then Equation (38) is equal to

$$\int_0^{\left(b - \frac{g_k B_{U_t}}{g_{U_t}}\right)} (b + B_{U_t})\,\mathrm{d}t = (b + B_{U_t})\left(b - \frac{g_k B_{U_t}}{g_{U_t}}\right) = b^2 + \left(1 - \frac{g_k}{g_{U_t}}\right) bB_{U_t} - \frac{g_k}{g_{U_t}} B_{U_t}^2.$$

If $b - \frac{g_k B_{U_t}}{g_{U_t}} \leq 0$ then Equation (38) is equal to 0. We thus obtain

$$\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[t < b - \frac{g_k B_{U_t}}{g_{U_t}}\right](b + B_{U_t})\right]\mathrm{d}t = \mathbb{E}\left[b^2 + (B_{U_t} - \min\{b, \frac{g_k}{g_{U_t}} B_{U_t}\})b - \min\{b, \frac{g_k}{g_{U_t}} B_{U_t}\}B_{U_t}\right]. \tag{39}$$

Fifth, we focus on the calculation corresponding to the fifth term of (13), that is, the case when $0 < t < b$ and $\frac{b-t}{g_k} \leq \frac{B_{U_t}}{g_{U_t}}$. We have

$$\begin{aligned}
&\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{U_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\right]\mathrm{d}t \\
&= \mathbb{E}\left[\int_0^\infty \mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{U_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\mathrm{d}t\right]
\end{aligned}$$

as we make use of Tonelli's Theorem. It follows that

$$\begin{aligned}
&\int_0^\infty \mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{u_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\mathrm{d}t = \int_{\left(b - \frac{g_k B_{U_t}}{g_{U_t}}\right)^+}^b \left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\mathrm{d}t \\
&= \left[-\frac{t^2}{2}\frac{g_{U_t}}{g_k} + tb\frac{g_k + g_{U_t}}{g_k}\right]_{\left(b - \frac{g_k B_{U_t}}{g_{U_t}}\right)^+}^b.
\end{aligned} \tag{40}$$

If $b - \frac{g_k B_{U_t}}{g_{U_t}} > 0$ then Equation (40) is equal to

$$\int_0^\infty \mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{U_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\mathrm{d}t = \left[-\frac{t^2}{2}\frac{g_{U_t}}{g_k} + tb\frac{g_k + g_{U_t}}{g_k}\right]_{b-\frac{g_k B_{U_t}}{g_{U_t}}}^{b} = \frac{g_k}{g_{U_t}}B_{U_t}(b + \frac{B_{U_t}}{2}).$$

If $b - \frac{g_k B_{U_t}}{g_{U_t}} \leq 0$ then Equation (40) is equal to

$$\int_0^\infty \mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{U_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\mathrm{d}t = \left[-\frac{t^2}{2}\frac{g_{U_t}}{g_k} + tb\frac{g_k + g_{U_t}}{g_k}\right]_0^{b} = b^2\left(\frac{g_{U_t}}{2g_k} + 1\right).$$

We thus obtain

$$\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[b - \frac{g_k B_{U_t}}{g_{U_t}} \leq t < b\right]\left(-t\frac{g_{U_t}}{g_k} + b\frac{g_k + g_{U_t}}{g_k}\right)\right]\mathrm{d}t = \mathbb{E}\left[\frac{g_k}{g_{U_t}}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\}\left(\frac{1}{2}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\} + b\right)\right]. \tag{41}$$

Sixth, we focus on the subtraction between the sixth term of (13), that is, the case when $0 < b < t$, and $\mathbb{E}\left[S_k(0,b)\middle|A = 0\right] = b$. We then have

$$\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[b < t\right]b - b\right]\mathrm{d}t = -\mathbb{E}\left[\int_0^\infty \mathbf{1}\left[0 < t < b\right]b\mathrm{d}t\right]$$

as we make use of Tonelli's Theorem. It follows that $-\int_0^\infty \mathbf{1}\left[0 < t < b\right]b\mathrm{d}t = -\int_0^b b\mathrm{d}t = -b^2$.

We thus have

$$\int_0^\infty \mathbb{E}\left[\mathbf{1}\left[b < t\right]b - b\right]\mathrm{d}t = -b^2. \tag{42}$$

In conclusion, summing Equations (34), (36), (37), (39), (41) and (42) we obtain

$$\begin{aligned}
\overline{S}_k^{(1)}(0,b) &= \int_{\mathbb{R}}\left(\mathbb{E}\left[S_k(0,b)\middle|A(-\infty,\infty) = 1, \tau = t\right] - \mathbb{E}\left[S_k(0,b)\middle|A(-\infty,\infty) = 0\right]\right)\mathrm{d}t \\
&= \mathbb{E}\left[\frac{1}{2}\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 + b\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} + \frac{g_k + g_{U_t}}{g_k}b\left(B_{U_t} - \min\{B_{U_t}, \frac{g_{U_t}}{g_k}b\}\right) - bB_{U_t}\right. \\
&\quad + b^2 + \left(B_{U_t} - \frac{g_k}{g_{U_t}}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\}\right)b - \frac{g_k}{g_{U_t}}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\}B_{U_t} \\
&\quad \left. + \frac{g_k}{g_{U_t}}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\}\left(\frac{1}{2}\min\{\frac{g_{U_t}}{g_k}b, B_{U_t}\} + b\right) - b^2\right] \\
&= \mathbb{E}\left[\frac{1}{2}\left(1 + \frac{g_k}{g_{U_t}}\right)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 + \left(b - \frac{g_k + g_{U_t}}{g_k}b - \frac{g_k}{g_{U_t}}b - \frac{g_k}{g_{U_t}}B_{U_t} + \frac{g_k}{g_{U_t}}b\right)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}\right. \\
&\quad \left. + \frac{g_k + g_{U_t}}{g_k}bB_{U_t} - bB_{U_t} + b^2 + bB_{U_t} - b^2\right] \\
&= \mathbb{E}\left[\frac{1}{2}\left(1 + \frac{g_k}{g_{U_t}}\right)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 - \left(b\frac{g_{U_t}}{g_k} + \frac{g_k}{g_{U_t}}B_{U_t}\right)\min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} + \frac{g_k + g_{U_t}}{g_k}bB_{U_t}\right].
\end{aligned}$$

Appendix C: Alternative way to derive light-traffic derivatives

As explained at the beginning of Section 3, for λ small enough, $\overline{S}_k(\lambda, b)$ can be approximated by a polynomial

$$\sum_{m=0}^{\infty} \lambda^m r_m(b), \tag{43}$$

where $r_m(b) = \dfrac{\overline{S}_k^{(m)}(0,b)}{m!}, m = 0, 1, 2, ...$

In this paper we used the approach initiated by Reiman and Simon [Reiman and Simon 1989] in order to derive expressions for the coefficients $r_m(b)$, see Appendix A. In this section we show how the coefficients $r_m(b)$ could alternatively have been obtained for the DPS model. We do this by using the integro-differential equation for the mean conditional sojourn time as obtained by Fayolle et al., see Equation (1). Before continuing, we like to emphasise that the calculations below are only valid for the DPS model, while the approach as taken in the paper is constructive and can easily be adapted to other models.

For the zeroth coefficient we have from Equation (43) $\dfrac{d\overline{S}_k(0,b)}{db} = \dfrac{dr_0(b)}{db}$ and from Equation (1) $\dfrac{d\overline{S}_k(0,b)}{db} = 1$. This immediately gives us

$$r_0(b) = b. \tag{44}$$

Since we assumed that for λ close to zero the function $\overline{S}_k(\lambda, b)$ can be approximated by $\sum_{m=0}^{\infty} \lambda^m r_m(b)$, we have for $m = 1, 2, \ldots$

$$\begin{aligned} r_m(b) = \lim_{\lambda \to 0} \frac{1}{\lambda^m} \left(\overline{S}_k(\lambda, b) - \sum_{i=0}^{m-1} \lambda^i r_i(b) - \sum_{i=m+1}^{\infty} \lambda^i r_i(b) \right) &= \lim_{\lambda \to 0} \frac{1}{\lambda^m} \left(\overline{S}_k(\lambda, b) - \sum_{i=0}^{m-1} \lambda^i r_i(b) \right) \\ &= \lim_{\lambda \to 0} \frac{1}{\lambda^m} \int_0^b \left(\frac{\partial \overline{S}_k(\lambda, \tilde{b})}{\partial \tilde{b}} - \sum_{i=0}^{m-1} \lambda^i \frac{dr_i(\tilde{b})}{d\tilde{b}} \right) d\tilde{b}. \end{aligned}$$

Now, substituting Equation (1) in the above formula, one can easily derive the terms $r_m(b), m = 1, \ldots,$ recursively. For the first and second order derivatives we obtain in this way

$$r_1(b) = \sum_{j=1}^{K} \alpha_j \frac{g_j}{g_k} \int_0^b \left(\mathbb{E}[B_j] + \left(\frac{g_k}{g_j} - 1 \right) \mathbb{E}[\min\{B_j, \frac{g_j}{g_k} \tilde{b}\}] \right) \mathrm{d}\tilde{b} \tag{45}$$

and

$$\begin{aligned} r_2(b) = \sum_{j=1}^{K} \alpha_j \frac{g_j}{g_k} \sum_{i=1}^{K} \alpha_i \frac{g_i}{g_k} \int_0^b \mathrm{d}\tilde{b} \Bigg(& \mathbb{E}[B_i] \left(\mathbb{E}[B_j] + \left(\frac{g_k}{g_j} - 1 \right) \mathbb{E}[\min\{B_j, \frac{g_j}{g_k} \tilde{b}\}] \right) \\ & + \left(\frac{g_k}{g_i} - 1 \right) \int_{\frac{g_j}{g_k}\tilde{b}}^{\infty} \mathbb{E}[\min\{B_i, \frac{g_i}{g_k}(x - \frac{g_j}{g_k}\tilde{b})\}] \cdot [1 - F_j(x)] \mathrm{d}x \\ & + \frac{g_k}{g_j} \left(\frac{g_k}{g_i} - 1 \right) \int_{\frac{g_j}{g_k}(\tilde{b}-b)}^{\frac{g_j}{g_k}\tilde{b}} \mathbb{E}[\min\{B_i, \frac{g_i}{g_k}(\tilde{b} - \frac{g_k}{g_j} z)\}] \cdot [1 - F_j(z)] \mathrm{d}z \Bigg). \end{aligned} \tag{46}$$

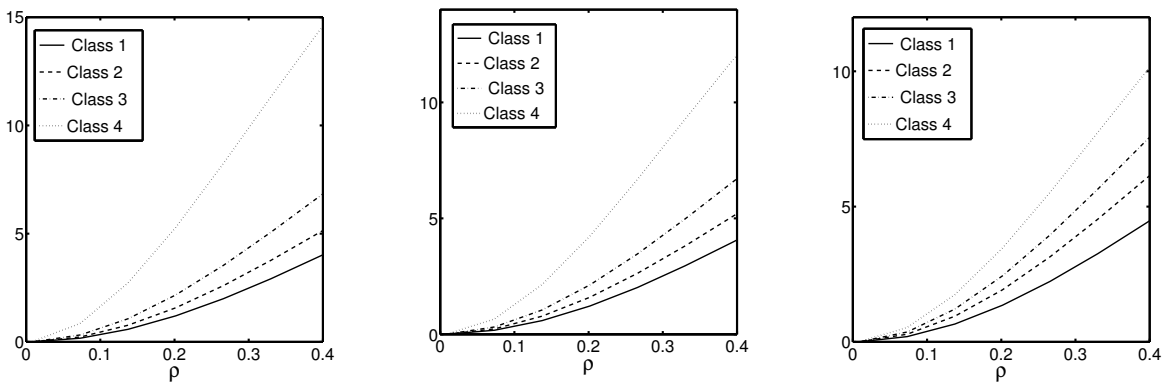

Fig. 14. Scenario 1: Difference of the relative errors of Equations (15) and (47) for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

We observe that $r_0(b)$ coincides with the zeroth light-traffic derivative obtained in Equation (12) and we verified that $r_1(b)$, obtained in Equation (45), coincides with the first light-traffic derivative shown in Equation (14).

We can now derive the following light-traffic approximation (of order 2) of the mean conditional sojourn time for a tagged class-k customer with service requirement b when λ is small

$$\overline{S}_k^{LT}(\lambda, b) = \sum_{m=0}^{2} \lambda^m r_m(b) = r_0(b) + r_1(b)\lambda + r_2(b)\lambda^2, \tag{47}$$

where $r_0(b), r_1(b)$ and $r_2(b)$ are given in Equations (44), (45) and (46), respectively. Using this result together with Proposition 4.1 and the heavy-traffic result (7) we obtain the third order light and heavy-traffic interpolation immediately

$$\overline{S}_k^{INT}(\lambda, b) = r_0(b) + r_1(b)\lambda + r_2(b)\lambda^2 + \frac{b}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2]/g_j} \frac{(\lambda \mathbb{E}[B])^3}{1 - \lambda \mathbb{E}[B]}. \tag{48}$$

We next assess the impact of having the second light-traffic derivative, thus obtaining a second order light-traffic approximation and a third order light and heavy-traffic interpolation. Then, we numerically compare the accuracy of Equations (15) and (47), and Equations (21) and (48).

In Figures 14 and 15 we plot the difference of the relative errors of Equations (15) and (47) for Scenario 1 and Scenario 4, respectively. Since we know that all relative errors are positive and since the resulting functions are also positive, this implies that the higher order light-traffic approximation is always more accurate. In Figures 16 and 17 we plot the relative error of the mean conditional sojourn time for Scenario 1 and Scenario 4, respectively. If we compare Figures 16 and 17 with Figures 3 and 6, we conclude that the accuracy of the interpolation for intermediate loads does not necessarily improve as the degree of the interpolation increases. In both cases, having a third order light and heavy-traffic interpolation reduces the largest relative error only for the case $\mathbb{P}(B_i \leq b_i) = 0.99$, while for $\mathbb{P}(B_i \leq b_i) = 0.01$ and $\mathbb{P}(B_i \leq b_i) = 0.5$ the third order approximation is worse than the second order approximation.

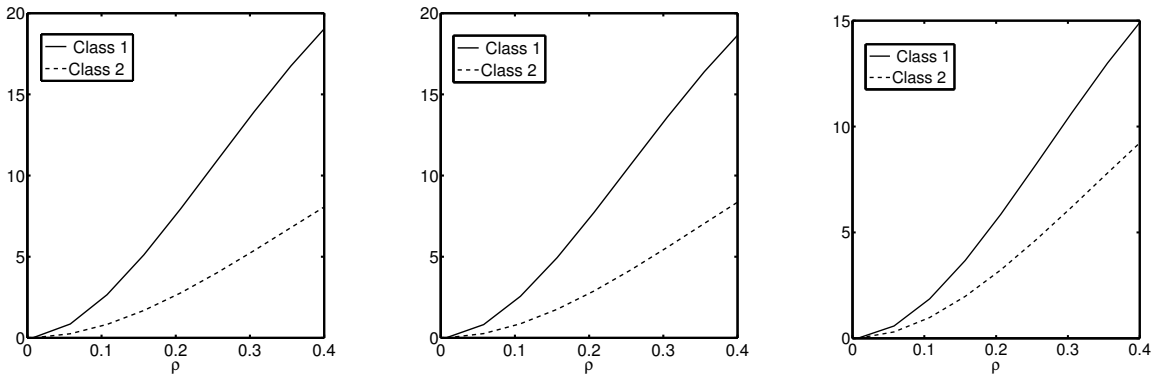

Fig. 15. Scenario 4: Difference of the relative errors of Equations (15) and (47) for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

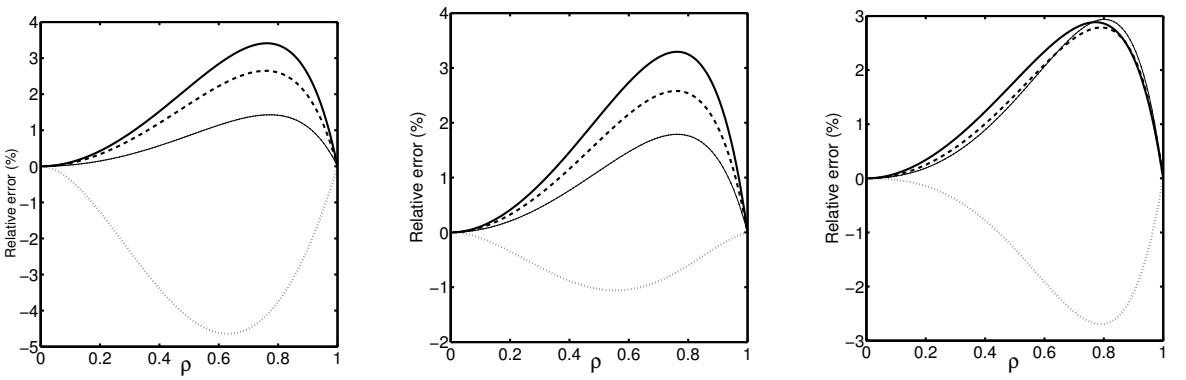

Fig. 16. Scenario 1: Relative error for the 3rd order mean conditional sojourn time for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

Appendix D: Proof of Corollary 5.1

For exponential service requirements we have the following equalities:

$$\mathbb{E}\big[\min\{B_j, b\frac{g_j}{g_k}\}\big] = \int_0^\infty \min\{b_j, b\frac{g_j}{g_k}\}\mathrm{d}F_j(b_j) = \int_0^{b\frac{g_j}{g_k}} b_j \mathrm{d}F_j(b_j) + \int_{b\frac{g_j}{g_k}}^\infty \left(b\frac{g_j}{g_k}\right)\mathrm{d}F_j(b_j) = \frac{1}{\mu_j}\cdot F_j\left(b\frac{g_j}{g_k}\right),$$

$$\begin{aligned}\mathbb{E}\big[\min\{B_j, b\frac{g_j}{g_k}\}^2\big] &= \int_0^\infty \min\{b_j, b\frac{g_j}{g_k}\}^2\mathrm{d}F_j(b_j) = \int_0^{b\frac{g_j}{g_k}} b_j^2\mathrm{d}F_j(b_j) + \int_{b\frac{g_j}{g_k}}^\infty \left(b\frac{g_j}{g_k}\right)^2\mathrm{d}F_j(b_j)\\ &= \frac{2}{\mu_j}\left(-b\frac{g_j}{g_k} + \left(b\frac{g_j}{g_k} + \frac{1}{\mu_j}\right)F_j\left(b\frac{g_j}{g_k}\right)\right),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[B_j\min\{B_j, b\frac{g_j}{g_k}\}\big] &= \int_0^\infty b_j\min\{b_j, b\frac{g_j}{g_k}\}\mathrm{d}F_j(b_j) = \int_0^{b\frac{g_j}{g_k}} b_j^2\mathrm{d}F_j(b_j) + b\frac{g_j}{g_k}\int_{b\frac{g_j}{g_k}}^\infty b_j\mathrm{d}F_j(b_j)\\ &= -\left(b\frac{g_j}{g_k}\right)^2\left(1 - F_j\left(b\frac{g_j}{g_k}\right)\right) + b\frac{g_j}{g_k}\frac{1}{\mu_j} + \left(1 - b\frac{g_j}{g_k}\frac{\mu_j}{2}\right)\mathbb{E}\big[\min\{B_j, b\frac{g_j}{g_k}\}^2\big].\end{aligned}$$

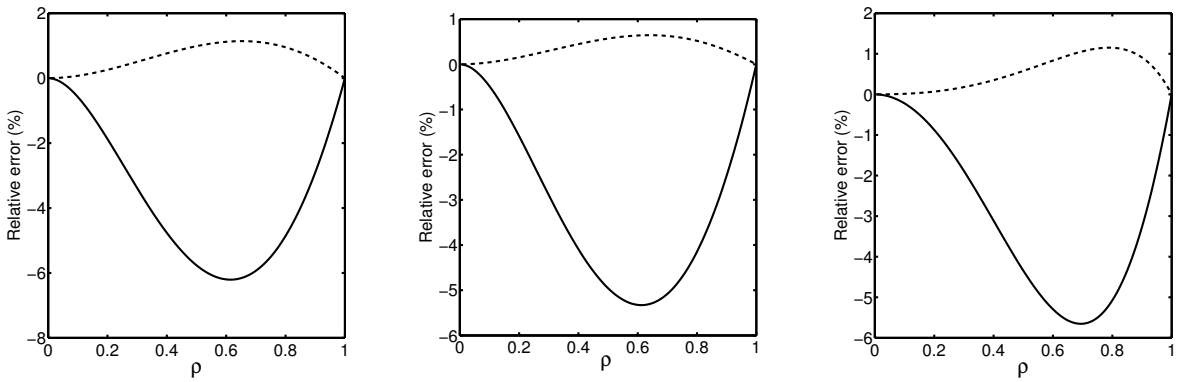

Fig. 17. Scenario 4: Relative error for the 3rd order mean conditional sojourn time for a tagged class-i customer with service requirement b_i such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left), $\mathbb{P}(B_i \leq b_i) = 0.50$ (middle), $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).

Then, considering Equation (21) and unconditioning on U_t we obtain

$$\begin{aligned}
&\overline{S}_k^{INT}(\lambda, b) \\
&= b + \lambda b \mathbb{E}[B] + \lambda \sum_{j=1}^{K} \alpha_j \mathbb{E}\left[\frac{1}{2}\big(1 + \frac{g_k}{g_j}\big) \min\{B_j, b\frac{g_j}{g_k}\}^2 - \big(b\frac{g_j}{g_k} + \frac{g_k}{g_j} B_j\big) \min\{B_j, b\frac{g_j}{g_k}\} + b\frac{g_j}{g_k} B_j\right] \\
&\quad + \frac{(\lambda \mathbb{E}[B])^2}{(1 - \lambda \mathbb{E}[B])} \frac{b}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2]/g_j}.
\end{aligned}$$

Together with the equalities given above, we then obtain (23).

The mean unconditional sojourn time, Equation (24), follows from Equation (23) together with

$$\sum_{j=1}^{K} \frac{\alpha_j}{\mu_j^2}\big(1 - \frac{g_k}{g_j}\big) \int_0^\infty \Big(1 - \mathrm{e}^{-\mu_j b \frac{g_j}{g_k}}\Big) \mu_k \mathrm{e}^{-\mu_k b} \mathrm{d}b = \sum_{j=1}^{K} \frac{\alpha_j}{\mu_j} \frac{(g_j - g_k)}{g_j \mu_j + g_k \mu_k}.$$